Nonribosomal Peptide Identification with Tandem Mass Spectrometry by Searching Structural Database

by

Lian Yang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Lian Yang

Abstract

Nonribosomal peptides (NRP) are highlighted in pharmacological studies as novel NRPs are often promising substances for new drug development. To effectively discover novel NRPs from microbial fermentations, a crucial step is to identify known NRPs in an early stage and exclude them from further investigation. This so-called dereplication step ensures the scarce resource is only spent on the novel NRPs in the following up experiments. Tandem mass spectrometry has been routinely used for NRP dereplication. However, few bioinformatics tools have been developed to computationally identify NRP compounds from mass spectra, while manual identification is currently the roadblock hindering the throughput of novel NRP discovery.

In this thesis, we review the nature of nonribosomal peptides and investigate the challenges in computationally solving the identification problem. After that, iSNAP software is proposed as an automated and high throughput solution for tandem mass spectrometry based NRP identification. The algorithm has been evolved from the traditional database search approach for identifying sequential peptides, to one that is competent at handling complicated NRP structures. It is designed to be capable of identifying mixtures of NRP compounds from LC-MS/MS of complex extract, and also finding structural analogs which differ from an identified known NRP compound with one monomer. Combined with an in-house NRP structural database of 1107 compounds, iSNAP is tested to be an effective tool for mass spectrometry based NRP identification.

The software is available as a web service at http://monod.uwaterloo.ca/isnap for the research community.

Acknowledgements

I would first like to thank my supervisor, Dr. Bin Ma, for directing the NRP database search project and providing insightful advices throughout my graduate study. He guided me from the very beginning, the fundamental of mass spectrometry, to the state-of-the-art algorithms in bioinformatics. I feel privileged to have such a great advisor.

I would like to thank my research collaborators in McMaster University, Dr. Nathan Magarvey, for visioning the utilization of informatics in natural product discovery and ensuring the project to be useful in practical lab research from a biochemist's perspective; Ashraf Ibraham, for compiling an in-house NRP database from literature, the lab work in growing bacteria and generating mass spectra from them. All of these facilitated the development and evaluation of the algorithms.

I would like to thank my thesis readers, Dr. Ming Li and Dr. Brendan McConkey, for generously spending time in reviewing my thesis and providing critical comments.

I would like to thank my parents, for encouraging me to pursue knowledge in science and providing the best possible education when I was a child. I would not be myself without their support and effort. I would like to thank my girlfriend, Ling Zhong, for the love and care she gave me.

At last, I would like express my gratefulness to my colleagues in the bioinformatics research group, Lin He, Xi Han, Daniel Dexter, Xuefeng Cui and Guangyu Feng, for all the inspiring discussions and the help they gave me.

Dedication

The thesis is dedicated to my parents for the encouragement and unconditional love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Nature has the power to craft an almost infinite number of unique and effective molecules. Among natural molecules produced by living organism, many have been discovered with a diverse sphere of bioactivities, such as antibiotics, immunosuppressants, toxins, and etc.[1]. These are certainly interesting properties that suggest pharmaceutical potentials. Over decades, bioactive natural products are highlighted in pharmaceutical studies as they are promising substances for drug development[2], and are also utilized to understand biochemical mechanisms[3]. According a review article published in Science Magazine in 2009 [4], approximately 50% of all new drug approvals in the past 25 years either come from natural products and their variants, or are semi-synthetics, which are synthesized using natural compounds as starting materials. (Figure 1.1)

Within the category of natural products, nonribosomal peptides (NRP) are a family of small molecules that are naturally produced as secondary metabolites by microorganisms. Optimized through natural selection, many nonribosomal peptides have been evolved to fit specific functions in certain biological processes. As secondary metabolites, nonribosomal peptides are not directly necessary in an organism's growth cycle, but often have a vital impact on the organism's continuing existence in adverse situations and play important roles in interspecies defense.

As such, nonribosomal peptides are always targeted in the frontier of the search for therapeutic agents. In the battle with infection diseases, the demand for powerful antibiotics is substantial. Vancomycin[5] (Figure 1.2), which is a nonribosomal peptide, was

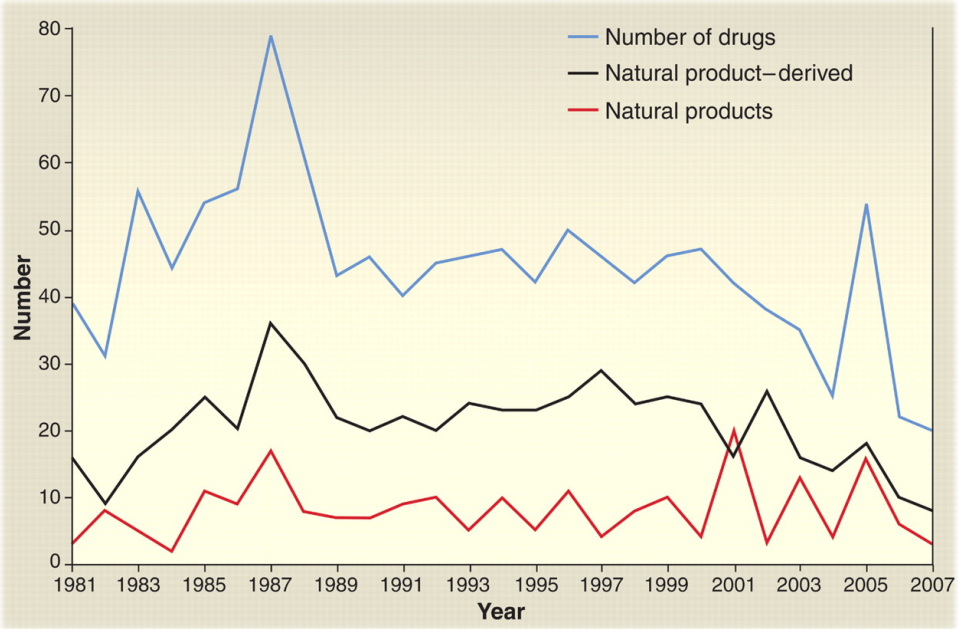


Figure 1.1: Number of drugs approved in the United States from 1981 to 2007. [4]

discovered and isolated in 1953. It was used but no longer capable as our last line of defense, as bacteria and pathogenic organisms are continuously evolving and building up resistance. Under the threat of "super bugs" infection, new drugs have been developed, including daptomycin[6], which originated from nonribosomal peptides. The need for advancing antibiotics discovery is real and urgent, and nonribosomal peptides are naturally a promising pool of candidates.

The range and scope of drugs that have been developed from nonribosomal peptides are also vast. In the antibiotics category, besides vancomycin, more NRP compounds have been developed to be approved drugs, such as gramicidins[7] and polymyxins[8]. Other than antibiotics, cyclosporine-A[9], rapamycin[10], are isolated to be an immunosuppressant drug, widely used in organ transplant to reduce the risk of rejection. As cytostatics, bleomycin[11], epothilone[12] and related analog compounds are developed to be anticancer agents in clinical treatment of various type of cancers. Figure 1.2 shows the molecular structures of some of the NRPs mentioned above. Overall, approximately 5% of 205 grouped families of nonribosomal peptides have become clinically approved drugs[13].

Beyond those nonribosomal peptides which are already known, more research efforts are spent in finding novel nonribosomal originated compounds with bioactivity[14]. Both

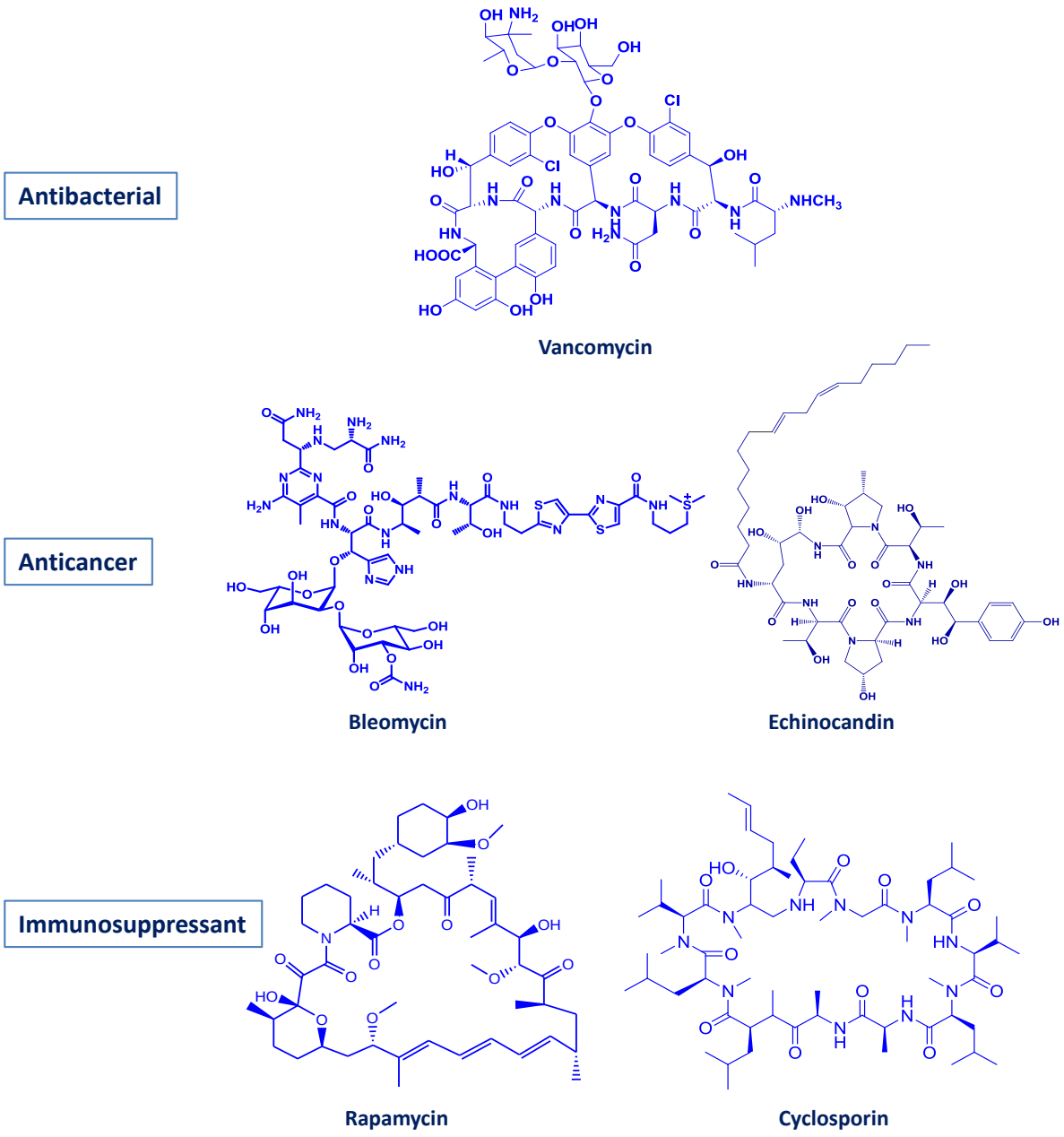


Figure 1.2: Samples structures of nonribosomal peptides as approved drugs.

new NRP structures and novel NRP analogs are targeted, as they could possibly be candidates or leads for new drug development. As for the methods, researchers tend to choose newly found microbial strains for fermentation screening, as they are more likely to produce unknown natural products. In a typical screening, microbial strains are cultured under various conditions, and often stressed with heat or ethanol shock to hopefully trigger the expression of biosynthetic genes that encoded for nonribosomal peptide synthetases (NRPS). Fermentations are screened for bioactivity with liquid chromatograph (LC), and tandem mass spectrometry (MS/MS) is then applied to interrogate the structural information of compounds in bioactive fractions. In the event that promising compounds exist, such compounds are then purified to allow structural confirmation with nuclear magnetic resonance spectroscopy (NMR).

Unfortunately, nonribosomal peptide discovery increasingly results in reconfirmation of known compounds. This highlights the importance of identifying known NRP compounds in an early stage and excluded them from further studies. This process is also referred as *dereplication*[4]. Dereplication with mass spectrometry is consider to be a cost competitive practice, since it only requires small amounts of compound mixed within the fermentation extracts. Conversely, the stages after mass spectrometry are time consuming and expensive. For instance, structural confirmation with NMR consumes expensive reagents and also purified analyte[15], which is laborious to prepare, and the effort could be wasted if the compound turns out to be a known NRP.

However, manual NRP dereplication is laborious at this moment even with mass spectrometry. To manually identify NRP from tandem mass spectra, each spectrum is examined and compared with standard spectra of NRPs collected from literature. Such method is inefficient, because the number of spectra can be hundreds when dereplicating with LC-MS/MS data acquired from microbial extract. Moreover, given the fact that there is currently no census NRP spectral library as reference, experimental spectra are only compared to a limited number of standard NRP spectra. This either requires a good knowledge of anticipated NRPs[14] in order to selectively construct the set of standard spectra, or would simply miss a hit.

Nonribosomal dereplication is time consuming and currently the roadblock hindering the throughput of novel NRP discovery. An automated solution that confidently identifies nonribosomal peptides with tandem mass spectrometry will substantially relieve the problem.

1.2 Objectives

For more than a decade, computational methods have become popular for the analysis of mass spectrometry data. Multiple database search algorithms have been successfully developed for linear peptide identification with tandem mass spectrometry. These algorithms computationally link MS/MS data back to the corresponding peptides in a protein database, thus reveal the structural information. We believe the workflow could be similarly adopted to nonribosomal peptides (Figure 1.3). A NRP database search algorithm is need to fill in the gap, which links MS/MS spectra of nonribosomal peptides to the correct structure in an NRP database.

In this research, the primary objective is to invent a database search algorithm that practically solves the dereplication problem for nonribosomal peptides, by computationally identifying NRPs from tandem mass spectra.

Specifically, the solution need to be practical with the following requirements satisfied:

- Usefulness
  - The solution should to be capable of identifying NRPs from either purified sample or within complex microbial extract at low abundance (μg/ml).
  - NRP identifications should be interpretable and associated with consistently calculated statistical scores. The statistical scores should truthfully help make the decision whether an NRP exists in the sample, thus advance the dereplication process.
  - A comprehensive database with a decent number of NRPs need to be provided along with the search algorithm.
- Correctness
  - NRPs with similar structure should be distinguishably identified by the algorithm. In other words, a spectrum must be precisely identified as the correct one when the database contains multiple NRPs of similar structures.
  - Identifications made by the algorithm should have a low false discovery rate. Tandem mass spectra of substances outside of the database should not be misidentified as one of the database NRPs.

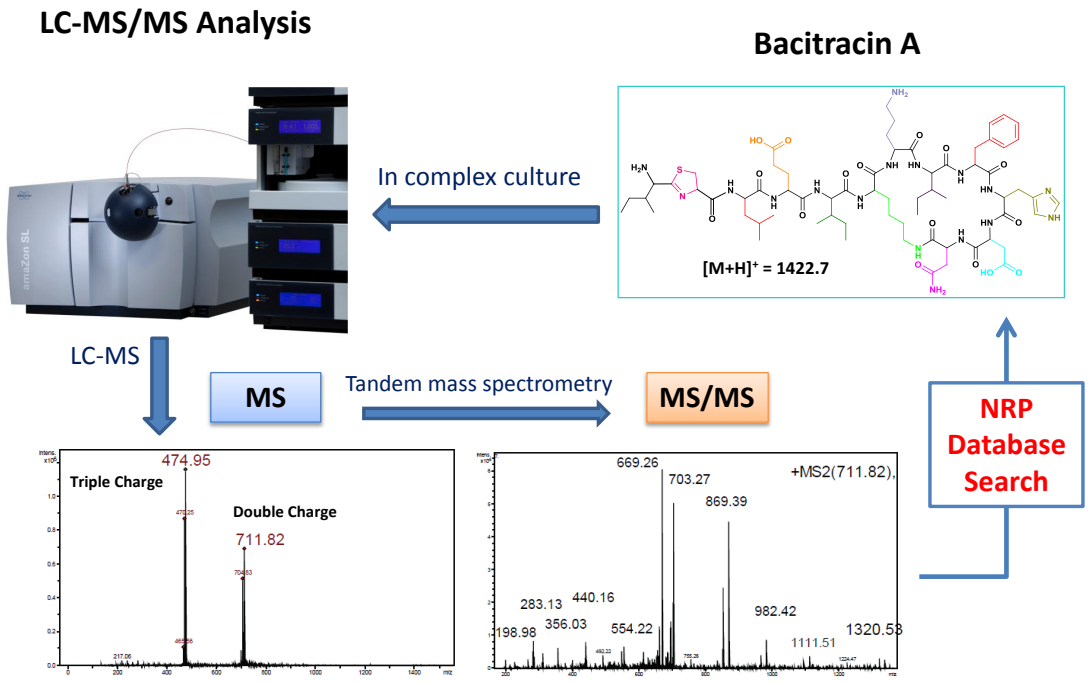


Figure 1.3: Workflow of MS/MS based nonribosomal peptides dereplication. Microbial fermentation with bioactivity is analyzed using an LC-MS system. With data dependent acquisition or pre-determined acquisition window, ionized compounds in the complex mixture are selected and then fragmented in tandem mass spectrometry. For each selected precursor ion, an MS/MS spectrum is generated with detected fragment ions. In such an experiment, the number of MS/MS spectra can be hundreds. Software is needed to compare those MS/MS spectra with a database of discovered NRPs, so that known NRPs in the fermentation can be identified and excluded from further studies.

It is common that a family of structurally similar NRPs can be produced simultaneously in the same microbial fermentation[16]. Tyrocidines, as an example, are only different at a few amino acid residues (Figure 1.4). When a known NRP is identified, it becomes more likely to find novel analogs coexisted in the same data set.

To take advantage of the given fact, we aim to expand the algorithm be an informatics tool for novel analog discovery. Precisely, we define analog as an NRP which is not in the database but only different at a small number of k building blocks to a database NRP, where a building block is the structural component between two adjacent amide bonds. As a secondary objective, we design the algorithm to be capable of identifying NRP analogs with $k = 1$. When it identifies an analog, it should indicate the original database NRP, from which the analog is derived, and also locate the position of the different building block.

Finally, the invented algorithms need to be implemented as software tools that can be easily used by wet-lab researchers to assist daily analysis of NRP deprelication.

1.3 Thesis Overview

The thesis is structured in the following chapters. In Chapter 2, we briefly review the nature of nonribosomal peptides, analyze their associated characteristics that prevent traditional identification algorithms from being applicable. The computational challenges for developing an NRP identification algorithm is then discussed. After that we review the pioneering works in the general topic of NRP identification. In Chapter 3, we first propose the design of iSNAP, the nonribosomal peptides database search algorithm, and demonstrate the underlying scoring scheme using bacitracin-A as an example. An experiment is then performed to determine the appropriate score thresholds. After that, the iSNAP algorithm is progressively challenged with a series of experiments, by which the algorithm's performance is analyzed. Chapter 4 illustrates the design and experiments for the analog search algorithm, which is proposed to be an expansion to iSNAP database search. In Chapter 5, we compile a list of conceivable projects that can be derived from this research, including new strategies for nonribosomal peptide identification and an innovative ways of utilizing the algorithm. Conclusions are presented in Chapter 6.

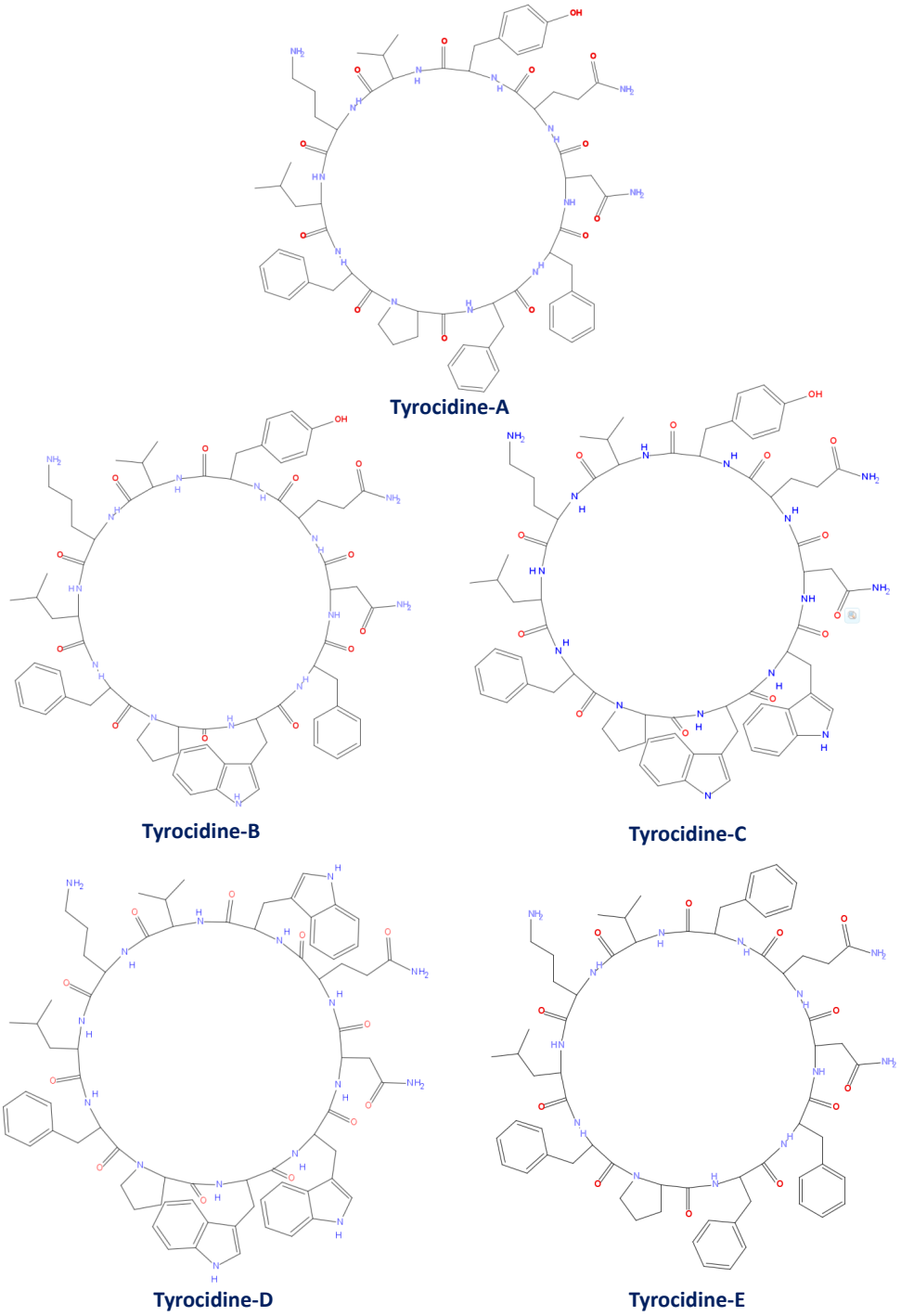


Figure 1.4: Molecular structures of tyrocidine A, B, C, D and E.

Chapter 2

Background

2.1 Nonribosomal Peptides

As the name suggests, nonribosomal peptides are a class of peptides that are not directly synthesized by ribosomes. In the synthesis of a ribosomal peptide, gene sequence is first transcribed to messenger RNA (mRNA), which serves as a template that carries the genetic information encoding the peptide. The template is then binded by ribosomes, which read the information and assemble the peptide. At ribosomes, transfer RNA (tRNA) recognizes codons from the mRNA and brings in the corresponding amino acid. Every codon is a short sequence consists of three nucleotides. A codon either encodes for one of the 20 proteinogenic amino acids, or functions as the indicator of the peptide terminal. The chain of translated amino acid are then assembled in ribosomes to form a linear peptide.

Nonribosomal peptides can be easily differentiated from its ribosomal counterpart.

Structurally, a ribosomal peptide is essentially a chain of amino acids linked by amide bonds. The amino acids, as a sequence, form a linear peptide backbone. However, nonribosomal peptides have much more diverse structures, often having a non-linear peptide backbone which is cyclic, branching, or a combination of the two[17]. As an example in Figure 2.1, bacitracin-A showcases the cyclization of amino acids. The C-terminal of peptide backbone connects with the side chain of a lysine residue and forms an amide bond. The cyclization leaves bacitracin-A a cyclic component as well a linear branch. More complex structures also exist, such as vancomycin, which has multiple cyclic components.

Nonribosomal peptides are also more complex in monomer composition. Ribosomal peptides are normally composed with 20 proteinogenic amino acids before post-translational

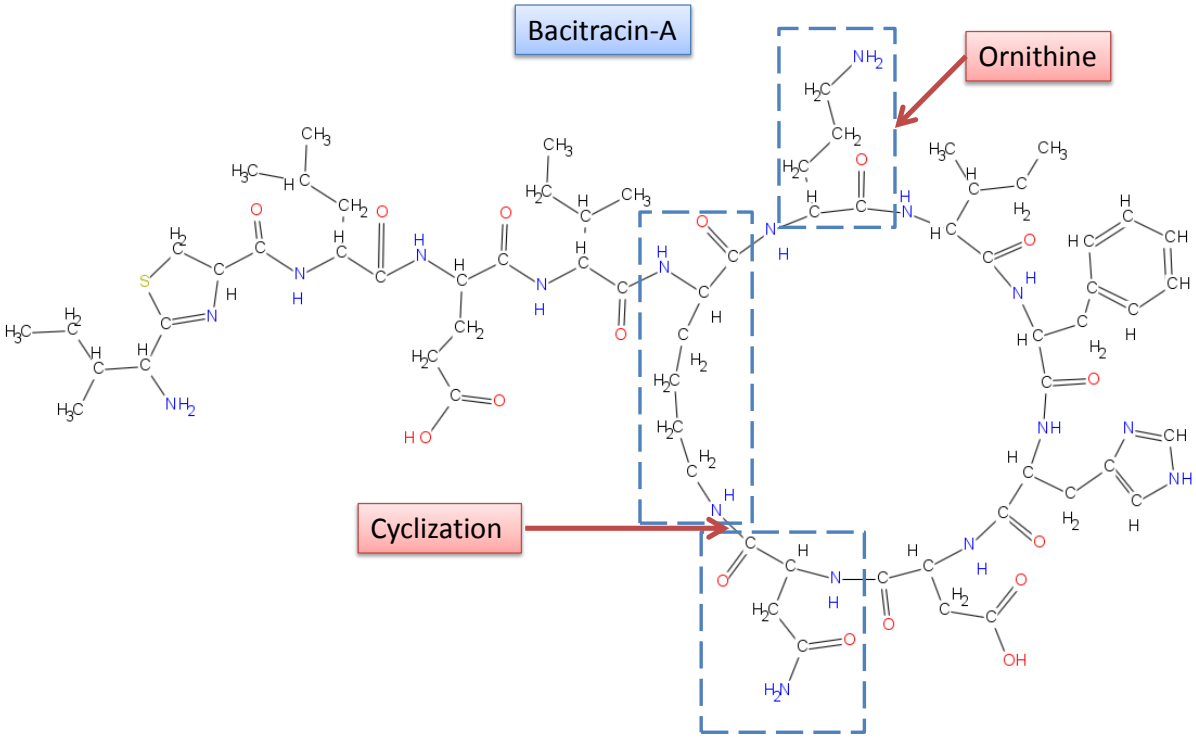


Figure 2.1: Molecular structure of bacitracin-A is distinct from that of ribosomal peptides. (1) Bacitracin-A contains an ornithine residue. Ornithine, as a non-proteinogenic amino acid, is not encoded by DNA. (2) Macrocyclization of the peptide backbone connects the C-terminal with the side chain of a lysine residue. The cyclization leaves bacitracin-A a cyclic component as well a linear "tail".

modifications. Nonribosomal peptides have many more types of monomers, which are not limited to proteinogenic amino acids. In NORINE database [13], more than 500 monomers in nonribosomal peptides have already been documented. On top of that, modifications are more commonly observed with nonribosomal peptides [18], which further increase the monomer diversity.

Back in 1960s, these peculiar features were observed in biochemical studies and implied that mechanisms other than ribosomal pathway is involved in the synthesis[19]. Experiments were designed to test the hypothesis, and it was confirmed with the observation that, even with the presence of RNAses or other inhibitors of ribosomal biosynthesis, natural production of theses compounds can still occur.

More comprehensive knowledge in synthesis of nonribosomal peptides was revealed with in-depth biochemical experiments in the 2000s[20]. The synthesis of nonribosomal peptides is directly facilitated by one or more nonribosomal peptide synthetases (NRPS), which are a type of specialized proteins. NRPSs are multi-enzyme complexes and vary in size, ranging from 10K Da to 1.6M Da. They are modularized in functionality. During the synthesis of nonribosomal peptides, each NRPS serves a particular role, specifically, as the initiation module, the elongation module or the termination module. It is noteworthy that NRPSs are not only the machinery for the synthesis, but also the template to determine the monomers in nonribosomal peptides. The synthesized linear peptide undergoes cyclization and further modifications before becoming the final product[18]. This pathway helps explain the incorporation of structural features in nonribosomal peptides.

It is noticed that NRPS is similar to polyketide synthetases (PKS) to a certain extent. NRPS and PKS can both be involved in a biosynthesis process, which resulting in compounds with mixed NRP and polyketide components[21]. This is observed in Epothilone[12], and is common in many secondary metabolites. Consequently, this further increases the structural complexity in nonribosomal peptides.

2.2 Computational Challenges for NRP identification

Before the investigation of challenges for NRP identification, computational methods for linear peptide identification is briefly reviewed. The quick review helps answer the question that, to what extent these traditional algorithms can inspire or resemble an algorithm for solving the problem of NRP identification.

For more than a decade, bioinformatics tools have become indispensable in mass spectrometry data analysis. Since the development of the SEQUEST algorithm in 1994 [22],

which is the first computational method to correlate tandem spectra and peptide sequences, multiple algorithms have been developed to identify linear peptides from tandem mass spectra, and some successfully commercialized. At present, the most popular software include Mascot[23], PEAKS[24], and X!Tandem[25], to name a few. With the capability to confidently identify a large number of peptides from complex proteomic data with decent sensitivity and accuracy, these software have been widely used in laboratories for scientific research.

Even though the goal is generally the same, different approaches are developed for linear peptide identification. Generally, algorithms can be characterized into three categories, database search, *de novo* sequencing, and spectral library search.

Database search is the most common approach. A database search algorithms identifies peptides by matching tandem spectra with hypothetical spectra of sequential peptides in a protein database. The hypothetical spectrum is generated by the algorithm on-the-fly, based on the amino acid sequence and fragmentation rules. As an example, with collision-induced dissociation[26, 27], which is a very common mechanism to fragment molecules in mass spectrometry, bond breakages mostly occur at amide bonds, resulting in b-ions and y-ions as the most abundant fragment ions (See Figure 2.2). Other fragment ions, such as a-ions and "satellite" ions with further loss of NH_3 or H_2O, are often observed as well. Following these rules, fragments can be generated *in-silico*, and the collection of mass-to-charge (m/z) ratios of these fragments become a hypothetical spectrum. The algorithm generate scores that indicates how well a tandem spectrum is explained by a hypothetical spectrum. Generally, when a good number of significant peaks in a tandem spectrum are matched, the score is high, then an identification is made. SEQUEST, PEAKS, Mascot, X!Tandem are software tools using database search approach for peptide identification.

Database search has been proven to be successful with the capability to identify a large number of peptides from complex proteomic data. However, this method requires a protein database that defines the search space. Without further design, it cannot detect unexpected peptides, for example, those with modifications or mutations. It is also not applicable in the situation that there is no database for the analyzed sample.

De novo sequencing is an approach that directly analyzes peak transitions in the tandem spectrum, and derives a best matching peptide, in most cases, composed with 20 proteinogenic amino acids. Since the search space is not limited by any database, *de novo* sequencing can identify peptides with mutations, or those from unknown genomes. PEAKS is currently the one of the state-of-the-art software in peptide *de novo* sequencing[29] and is commercially available. Other *de novo* sequencing software includes Lutefisk[30] and PepNovo[31].

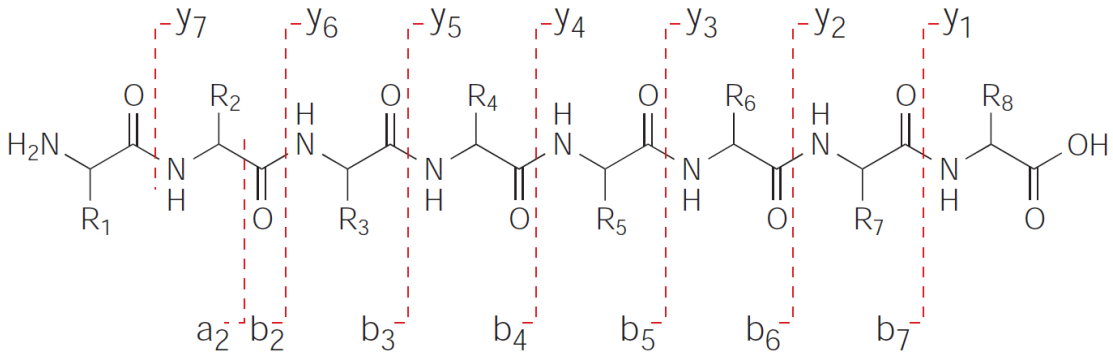


Figure 2.2: Typical fragmentation pattern of ribosomal peptides in collision-induced dissociation. [28]

Spectral library search is an alternative to database search. Instead of using hypothetical spectra generated from peptide sequence, tandem spectra are matched with a library of real spectra. The sequence of these library spectra are previously determined during the compilation of the library. Therefore, an identification is made when a tandem spectrum is matched by a library spectrum. Overall, library search is faster than database search, since it skips the step of *in-silico* fragmentation. It often generates identifications with better accuracy, because library spectra not only provides the realistic information on fragments' mass-to-charge ratios but also ion intensities. At this moment, the shortcoming of this approach is the limited coverage and availability of spectral libraries. However, library search is deemed as a promising approach, as comprehensive spectral libraries of peptides, such as PeptideAtlas [32], are being constructed with expanding coverage across human, mouse, yeast, and several other organisms.

For all three approaches, scoring is always an important component in a peptide identification algorithm. Scores are indicators about how well a spectrum represents a peptide. and they also facilitate ranking peptide candidates to find the best match for a spectrum. A scoring scheme distinguishes true identifications from random matches, and is the core component of any algorithm. Such a score can be directly calculated based on matched peaks. However, such a straightforward score is often not comparable across the identifications on different spectra. It is therefore more usable if the reported scores are calculated with statistical methods, so that a significance meaning can be established.

Unfortunately, these traditional algorithms are only designed to identify linear peptides, and not applicable to nonribosomal peptides which often have complex structures with cyclic and/or branching backbone. We assume that an NRP identification algorithm can be developed using at lease one of the above proven approaches, as suggested by similarity between the two problems. However, the following challenges must be addressed before an approach can be successfully adapted for NRP identification.

The first challenge is the data representation of structures, In traditional algorithms, linear peptides are represented as strings of amino acids. Each amino acid is stored as a single letter code. Using strings as the data structure, *in silico* fragmentation is a very natural process, in which a string is spitted sequentially at each position. Each time, two fragments are generated as the substrings from the cleavage site to either one of the string terminals. However, such a simple string is not adequate to represent a nonribosomal peptide with complex structures or having non-peptide components. We argue that an encoding system need be used, which preserves the structural information. It should be easy to store, and also friendly to computational analysis.

Secondly, other than the data structure, the algorithm itself is also challenged by NRP's structural complexity. For a database search algorithm, *in-silico* fragmentation is needed. As demonstrated in Figure 2.3, if the pure cyclic structure is only dissociated at one amide bond at a time, as it is done for linear peptides, the resulting fragments are going to be different versions of linearized peptide cycle, and have the exact same mass[33]. This is nearly useless in revealing structural information. Therefore, fragments generated with further dissociation on these linearize peptides, which also known as internal fragments, are more crucial for a successful identification of NRP structures.

As for *de novo* sequencing, algorithms have been developed with the assumption of linear structure. It would be a difficult problem to accurately reconstruct the complex structure of an NRP with sub-sequences derived from peak transitions in a spectrum. However, even sub-sequences are hard to be figured out as NRPs can be composed of more than 500 types of monomers, rather than the 20 proteinogenic amino acids. This drastically increases the search space, and make the traditional *de novo* sequencing algorithms not usable.

The third challenge concerns the availability of NRP database or NRP spectral library. A database or library indispensably defines the search space of the the algorithm. For spectral library search, although the matching algorithms are intrinsically applicable to any spectra, including spectra of NRPs, the approach is hindered by the absence of NRP spectral library. Currently, there is no NRP spectral libraries available. Though many labs have been generating and compiling spectral data from pure or semi-pure NRP compounds,

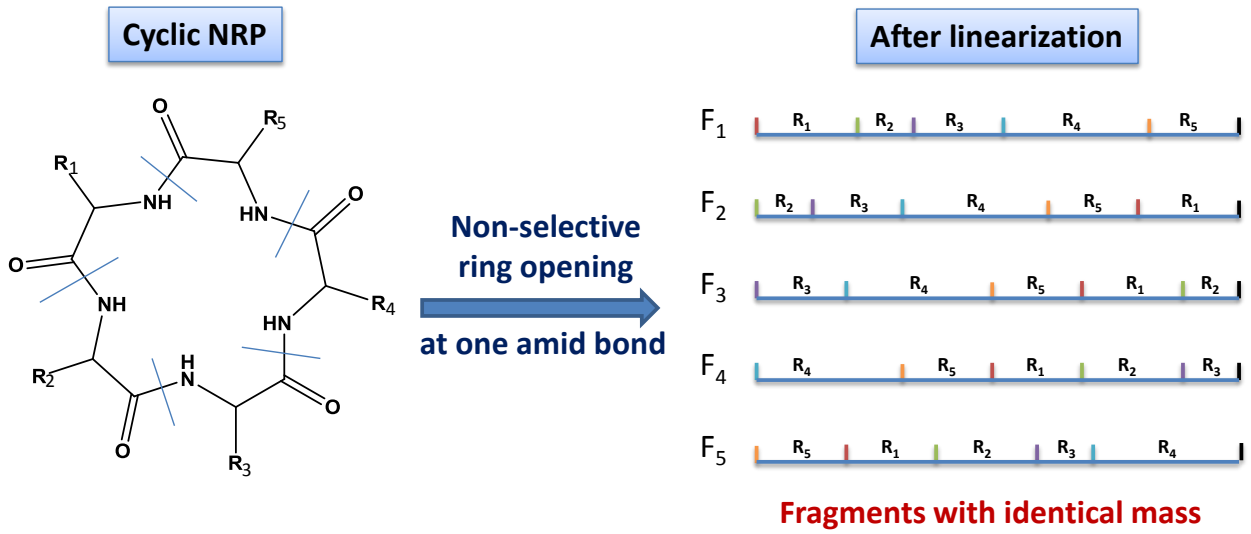


Figure 2.3: The figure demonstrates the linearization of a purely cyclic nonribosomal peptide. Using the fragmentation rule for ribosomal peptides, the cyclic structure is only dissociated at one amide bond at a time, leaving different versions of linearized fragments. These fragments have the same amino acid composition, but the amino acid residues are in assorted orders as the result of different cleavage site. With identical mass, the fragments cannot be differentiated by mass spectrometers, therefore cannot provide any structure information.

these spectral sets are usually small in size, have varying quality, and may not be available to the public.

Fortunately, data availability is much less an issue for structural database search, which only need to molecular structures of NRPs to search with. Norine [13], as a comprehensive database specialized for NRPs, has been freely available since 2008. Besides that, NRP structures can be collected from the small molecules databases in PubChem. Collecting NRP structures takes substantially less efforts than compiling a spectral library.

With the consideration of the above challenges, we conclude that database search is comparably the most feasible approach for solving the NRP identification problem. In the developing of an NRP database search algorithm, a workflow can be similarly established, with special efforts in addressing the challenges in data representation, *in-silico* fragmentation and database preparation. Moreover, a new robust scoring scheme is needed, which should efficiently distinguish true and false matches between experimental spectra and hypothetical spectra generated from NRP structures.

2.3 Related Work in NRP identification

In spite of the fact that NRP related informatics research is still in its infancy, there are already a number of pioneering works that laid the foundation for solving the NRP identification problem. We review some of these works in this section, and comment on their contributions and limitations.

2.3.1 Tandem Mass Spectra Interpretation of Cyclic NRPs

In 2009, Liu. etc [34] proposed the program MS-CPA, which annotates tandem mass spectra obtained with collision-induced dissociation from cyclic NRPs. The program takes an NRP structure and a spectrum as the input. It annotates the spectrum by highlighting peaks that are matched by the fragment ions derived from the structure. It is known that the complexity of NRP structures consequently increases the complexity of the resulting tandem mass spectrum. Knowing the structure, manually annotating a tandem spectrum can still be a difficult task. It was demonstrated that the program was capable of annotating seglitide and tyrocidines, and was used to confirm the sequence of two newly discovered NRPs, desmethoxymajusculamide-C and dudawalamide-A

In the development of MS-CPA, it was noticed that up to 15% of major peaks in a spectrum can not be explained by fragments generated from the corresponding NRP structure.

With further investigation, they realized that some of these unanticipated fragments could be explained by first scrambling the monomers in the cyclic peptide backbone before any known fragmentation rules were applied. It was truly unanticipated but was just a confirmation to the bizarre fragmentation behavior of NRPs. These unusual fragments were first described by Harrisons et al. in 2006 [35], as non-direct sequence (NDS), distinguished from direct sequence (DS) fragmented by traditional rules.

The program facilitates the elucidation of tandem spectra, and helps understand the fragmentation behavior of cyclic NRPs. However, the program is not a solution for NRP identification. The annotation requires an NRP structure as an input, assuming it is the correct one. No scoring is built on top of the annotation, thus we cannot use it to rank the annotations on a spectrum using different NRP as candidates. Besides, the program is limited to cyclic NRPs with a perfect cyclic backbone without any other complex components, thus, not applicable to the majority of NRPs.

2.3.2 *De Novo* Sequencing of Cyclic NRPs

Pioneering work by Ng et al. [36] has initially tackled the *de novo* sequencing problem for cyclic NRPs, which was considered much more challenging than database search. A *de novo* sequencing algorithm using MS3 spectrum was proposed and demonstrated to be successful in reconstructing cyclic NRPs. Although the algorithm only aimed to solve the problem for perfectly cyclic NRPs, the problem was still difficult. For a cyclic NRP, the amide bonds in the cycle are potentially opening sites. In the stage of MS2, with collision energy carefully controlled, each precursor ion of cyclic NRP is fragmented at only one amide bond. The breakage generates different linearized versions of the cyclic peptide[33], as illustrated in Figure 2.3, and no other fragments. These linearized versions have the same mass and undergo further fragmentation in MS3. Thus, the MS3 spectrum essentially appears as the superposition of internal fragment ions from each linearized peptide. The MS3 spectrum contains more structural information than the MS2, but it is difficult to interpret.

To sequence the cyclic NRP, the proposed algorithm first extracts a list of most prevalent peak transitions in the MS3 spectrum by using spectral auto-convolution. The technique was commonly used in signal processing for finding repeating patterns, such as periodic signals buried under noise, and was introduced to mass spectrometry analysis by Pevzner et al. in 2000 [37]. This step reveals the amino acid composition, as prevalent peak transitions are most likely corresponding to mass of monomers in the NRP. Then, a heuristic algorithm uses these mass values as the "alphabet" to align sub-sequences and

reconstruct the most probable NRP. In this way, the algorithm does not need to use a pre-defined monomer set, such as the 20 proteinogenic amino acids. Therefore, it avoids the problem caused by enormous types of monomers in NRPs.

The algorithm performed excellently in sequencing tyrocidines, cyclosporin and surfactin, and was marked as a significant step in bioinformatics research for NRPs. However, the *de novo* sequencing algorithm was specifically designed for purely cyclic NRPs, which are not applicable to the majority of NRPs. Moreover, the algorithm was only tested with purified samples, and the mass spectrometer was specifically optimized for the tested NRPs in order to generate clean fragmentation. It limits the algorithm's usability as a practical tool for NRP dereplication, where complex extract need to be processed in high throughput.

2.3.3 NORINE: A Database of Nonribosomal Peptides

NORINE is the first and most comprehensive database dedicated to nonribosomal peptides [13]. The NORINE project started in 2008 as a collaboration project between the SEQUOIA bioinformatics research group of LIFL (Laboratoire d'Informatique Fondamentale de Lille), INRIA (Institut National de Recherche en Informatique et en Automatique) and the NRPS team of ProBioGem laboratory. The objective is to provide a versatile platform that enables research in nonribosomal peptides. Such a platform includes a documented NRP database as well as tools for browsing. The NORINE database started with around 700 NRP entries, and is continuously growing ever since. Up to the year of 2010, the database already consists of 1122 NRPs, grouped into 205 families. Each NRP entry is comprehensively annotated with information of its structure, bioactivities, origins and so on. The NORINE website provides a set of tools, including structural pattern search which allows users to search for NRPs with a particular substructure.

NORINE database is a freely accessible to researchers. It sets a solid foundation for the development of database search algorithms. However, for unknown reasons, the NRP database has not been updated in the last two years. There are newly discovered NRP structures published in journals, such as the Journal of Antibiotics, but not included in NORINE.

Chapter 3

iSNAP for Nonribosomal Peptide Database Search

3.1 Method Overview

iSNAP is proposed as a bioinformatics tool to aid and assist in the dereplication of nonribosomal peptide (NRP) compounds from complex mixtures. The program has evolved from the traditional database search approaches, which are strictly applied to sequential peptides, to one capable of identifying complicated peptide structures. It has been designed to identify NRPs which may have cyclic and branching structures.

Figure 3.1 outlines the three components of iSNAP. The first is the structural NRP database. It contains 1107 nonribosomal peptides, and ultimately defines what can be possibly identified with the algorithm. For each collected compound, a list of possible fragment ions is generated and stored. We refer the list as the hypothetical spectrum, which is used to match with the input MS/MS spectra. The second component is the algorithm for NRP database search. A scoring scheme is designed to evaluate the matches between input MS/MS spectra and hypothetical spectra of NRPs in database. Significance scores are calculated for each match and enable the identification of NRPs. The third component is the algorithm for NRP analog search. This component is an extension to the database search algorithm. It utilizes the identified NRPs in database search as seeds, and re-analyzes the input MS/MS spectra to find analogs to a seed NRP with structural difference at one building block. The first two components, the database and the search algorithm, are described in this chapter, and the analog search is covered in Chapter 4.

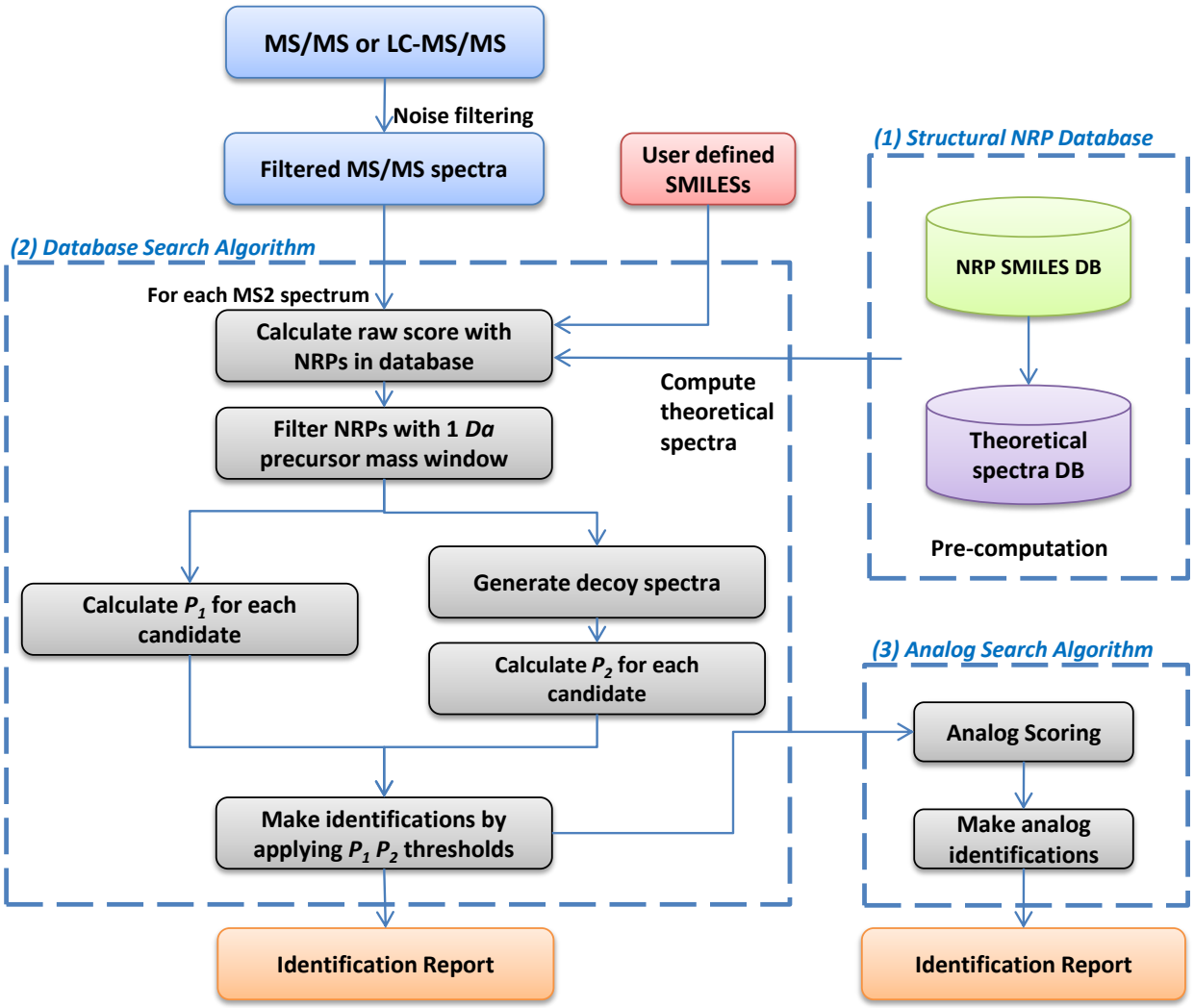


Figure 3.1: Workflow of NRP database search with iSNAP. As pre-computation, iSNAP generates hypothetical spectra of structures in the NRP database based on multiple cleavages at amide bonds. In the process of NRP identification, the noise peaks in input MS/MS spectra are first filtered. iSNAP analyzes each input MS/MS spectrum, either within LC-MS/MS acquired by *data dependent acquisition* or simply within a list of MS/MS scans. It scores the input MS/MS spectrum with the hypothetical spectrum of structures in the NRP database. Two scores, P_1 and P_2 are calculated for each input MS/MS spectrum, indicating the significance of the best matching NRP structure. If both scores are above their corresponding thresholds, the best matching structure is reported as an identified NRP and linked to the MS/MS spectrum.

3.2 Structural Database

3.2.1 Collection of Nonribosomal Peptides

A in-house database of nonribosomal peptides was collected by research collaborators from Nathan Magarvey Lab at McMaster University. In constructing this database, nonribosomal peptide were collected from the NORINE database[13], PubChem and the Journal of Antibiotics. Structures were first reconstructed in Chemdraw software and then converted to SMILES code[38]. SMILES stands for *Simplified Molecular Input Line Specification*, which is a standard encoding system that represents chemical molecules into a linear string while preserving all the structural information. The database is formatted in text files, where name and SMILES code of each compound are stored in pairs, and allows the algorithm to load efficiently.

The SMILES format is parsed by iSNAP and converted to a graph which stores atoms and bonds as vertices and edges, respectively. Represented by graph data structure, nonribosomal peptides can be annotated computationally by iSNAP, which identifies the linear, cyclic and cyclic-branching components. Furthermore, peptide backbone in a molecule is also annotated through amide bonds, providing linkage information of monomer blocks.

In total, the database consists of 1107 compounds, produced mostly by nonribosomal peptide synthetases. The mass distribution of database compounds is shown in Figure 3.2. Besides that, it is worth to mention that iSNAP allows researchers to upload a series of SMILES codes to be included into the database for dereplication or to fulfill their specific research purposes.

3.2.2 *In Silico* Fragmentation and Hypothetical Spectra

iSNAP processes the SMILES code of each NRPs to create the corresponding hypothetical spectrum. From the NRP's molecular structure, a unique series of fragments are created, based on fragmentations on their amide cleavage sites. These generated structural fragments are calculated estimations as how the protonated NRP may fragment in the collision-induced dissociation (CID)[26, 27] within the gas phase of an MS/MS experiment. To generate these fragments, iSNAP first loads the structure from its SMILES code and labels all the linear, cyclic and cyclic-branching components. Each amide bond is tagged as a potential site for cleavage. The hypothetical spectrum is compiled by exhaustively fragmenting the structure. The program enumerates every combination of two amide bonds, which are previously tagged, within the structure. For each combination, it

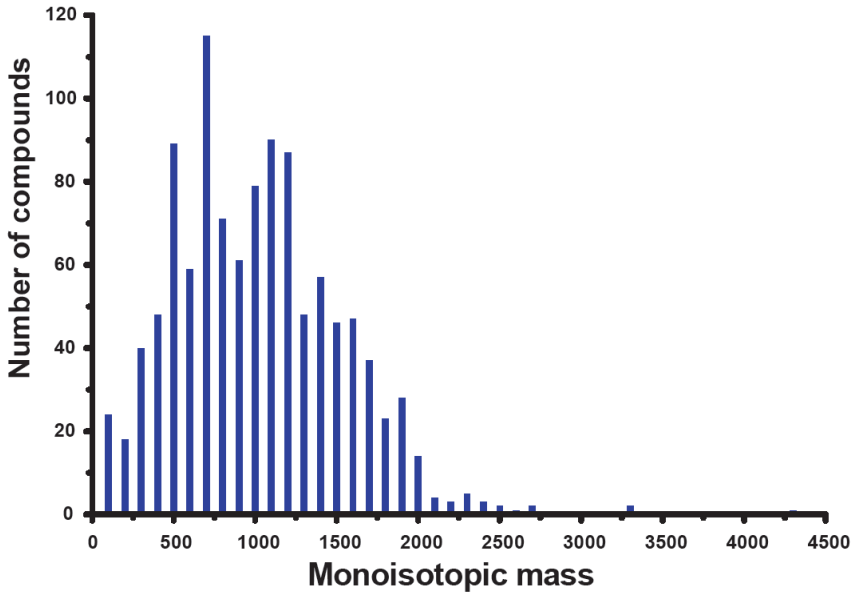
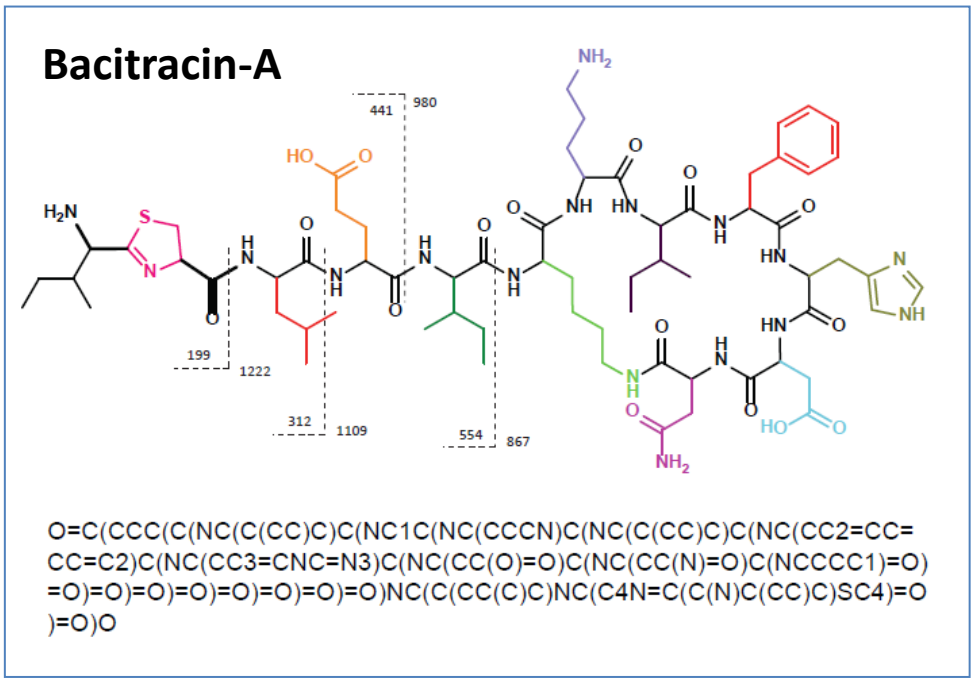


Figure 3.2: Mass distribution of the 1107 database NRPs.

breaks the structure at the two bonds, and collects the fragments. This process yields a set of generated structural fragments. As an example, Figure 3.3 shows the SMILES code of bacitracin-A and samples of generated structural fragments.

From the 1107 NRP structures, a library of 1107 hypothetical spectra was created with the above process. In a hypothetical spectrum, mass-to-charge ratios (m/z) are calculated using the fragments resulted from amide cleavages described above, assuming the precursor ion is singly charged. To generate the m/z values, mass offsets of +H and +H+1 are added for each fragment to account for protonation and the first isotope ion, respectively. Additional m/z values are added as well with further consideration of common neutral losses (water, ammonium and carbon monoxide).

The m/z values in constructing a hypothetical spectrum represents where a fragment ion is likely to be observed in a real singly-protonated MS/MS spectrum of the NRP. Thus, when the hypothetical and experimental spectra of the same NRP are compared, a significant number of high-intensity peaks should be shared by both spectra. However, if the two compared spectra are unrelated, while random matching can still produce some shared peaks, the number and intensities of the shared peaks is usually much less. Thus, by examining the shared peaks, information can be obtained about whether the experimental spectrum is from one of the known NRPs in the database. A hypothetical spectrum con-



Fragment #	SMILES	Molecular structure	Molecular weight
15	O=CC(NC(=O)C(N)CC1=CC=CC=C1)CC=2N=CNC=2		286.14
16	O=CC(NC(=O)C(NC(=O)C(N)CC(C)C)CCC(=O)O)C(C)CC		357.23
24	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)C(C)CC)CC1=CC=CC=C1)CC=2N=CNC=2)CC(=O)O)CC(=O)N		628.29
43	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)C(C)CC)CCCCN)CCCN)C(C)CC)CC1=CC=CC=C1)CC=2N=CNC=2)CC(=O)O)CC(=O)N		983.56

Figure 3.3: SMILES code and sample theoretical fragments of bacitracin-A.

tains multiple diagnostic fragments for the NRP and allows for the efficient dereplication. Input MS/MS data files, containing experimental spectra, can be searched and compared to the hypothetical spectra of database NRPs in order to determine the significance of the matched fragment peaks.

Compared to the calculation of hypothetical spectra for linear peptides in traditional database search algorithms, the calculation for NRPs is time consuming. It involves the parsing of SMILES code, tagging chemical components in graphs, and then sub-graphs can be generated as fragments. The runtime would be not acceptable if performing the calculation on-the-fly, as it was done traditionally. However, hypothetical spectra are static information which stay fixed each time a search is launched. Therefore, we move the calculation for database NRPs off-line, as a pre-computation. The generated hypothetical spectra are stored with the NRP database, each linked to the corresponding SMILES entry. In this way, it vastly speeds up the search algorithm, as only the hypothetical spectra of user-uploaded NRPs, instead of the whole NRP database, need to be calculated in a search.

3.3 Scoring Scheme

Having introduced how hypothetical spectral can be computational generated from compounds in the NRP database, the next step is to design a scoring scheme that serves two purposes. First, given the MS/MS spectrum of an analyte, the scoring scheme is used to retrieve the highest-scoring hypothetical spectrum. *Assuming the analyte's structure is in the library*, it should be the one that was used to compute the highest-scoring hypothetical spectrum. Secondly, the scoring scheme should help judge whether the assumption that the analyte is in the library is true.

Three scores are computed for these two purposes: (1) raw score, (2) P_1 score, and (3) P_2 score. Raw score measures the matching quality between an MS/MS spectrum and a hypothetical spectrum. It is used by the algorithm to rank the compounds in the database for a given spectrum. However, the absolute value of the raw score is not an interpretable measurement to judge whether the match is real or purely random. The P_1 and P_2 scores are designed to serve these purposes. Both P_1 score and P_2 score indicate the significance of a spectrum-compound match but in two different ways. They are calculated based on raw scores.

3.3.1 Raw Score

The raw score represents the basic spectrum-matching score, indicative of how well an MS/MS spectrum is matched with a hypothetic spectrum that is generated from a linear SMILES string. To calculate a raw score for a match, the algorithm first loads the singly charged or protonated hypothetical spectrum from the database. A simple noise filtering step is performed on the input MS/MS spectrum in order to remove peaks with low intensity values. Peaks with relative intensity less than 0.5% are removed, where the relative intensity of a peak is the intensity ratio between the peak and the highest peak in the spectrum. Such noise filtering is applied in order to reduce the possibility of randomly matched peaks, and is a common practice in designing of peptide-spectrum matching score in proteomics. The remaining peaks are then matched with the hypothetical spectrum.

As mentioned previously, hypothetical spectra are generated with the assumption that the precursor ion is only charged with one proton. However, an input MS/MS spectrum can be multiply charged. It is very common to see a precursor ion is doubly or triply charged. In such case, the algorithm adjusts the singly charged hypothetical spectrum to account for difference in charge states. The multiply charged ions are added by assuming additional protons are attached to the structural fragments. When the parent ion of the MS/MS spectrum has charge k, the m/z values of hypothetical fragments with charges up to k are combined to form the charge-k hypothetical spectrum.

By using a mass error tolerance of $0.1Da$, the algorithm finds all peaks in the input spectrum that are matched by the theoretical spectrum, and computes the Raw score as

$$Raw\ score = \sum_{each\ matched\ peak\ m_i} log_{10}(200 * relative\ intensity\ of\ m_i)$$

The factor 200 in the formula ensures that a peak with relative intensity 0.5% contributes score 0. The mass error tolerance $0.1Da$ is set empirically to accommodate the random and system errors arising from low resolution mass spectra. The mass tolerance value is, in essence, due to the mass accuracy of the mass spectrometer. Values set to low will limit the amount of matched fragments, while higher values will generate a greater number of matches, possibly increasing the chance for random assignments.

For each MS/MS spectrum, the raw score is calculated against the database compounds within a mass range of $0Da$ to $[M] + 100Da$, where $[M]$ represents the MS/MS spectrum's precursor mass. Having a relaxed mass range ensures sufficient raw scores are calculated for the reliable estimation of a statistical distribution, which is important for the calculation of the P_1 score as presented later. Meanwhile, the mass upper bound of $[M]+100Da$ avoids the

potential bias arising from large molecular weight compounds, as more structural matches could be made randomly for larger compounds. Only database compounds within the mass range of $[M] \pm 1Da$ are subjected to P_1, P_2 calculation, and are considered as candidates for that MS/MS spectrum.

3.3.2 P_1 Score

A P_1 score is introduced as a normalized version of the raw score in order to add a statistical significance. To calculate the P_1 score, the input MS/MS spectrum is scored against the hypothetical spectra of all database NRPs within the $0Da$ to $[M]+100Da$ mass range, the statistical distribution of the raw scores closely matches a gamma distribution. The gamma distribution is selected empirically as it best fits the data. The parameters required by the gamma distribution are estimated with the maximum-likelihood method, and the estimated gamma distribution curve is plotted to the raw scores. (See Figure 3.4)

For each compound, the p-value is the exceedance frequency at the compound's raw score, which is the area under the curve and to the right of the Raw score. The p-value represents the probability of that a random structure is scored higher with the MS/MS spectrum than the current structure. A low p-value indicates the match is unlikely random and therefore is likely a correct one. A P_1 score is derived from the p-value, whereby, a higher value represents a greater significance of being correctly matched.

$$P_1 \; score = -10 \, log_{10}(p\text{-value})$$

3.3.3 P_2 Score

Similarly to the P_1 score, a P_2 score is calculated for each candidate structure within the mass range of $[M] \pm 1Da$.

The P_2 score measures the matching quality between an MS/MS spectrum and the hypothetical spectrum of an NRP compound. Decoy spectra are generated by shifting the original MS/MS spectrum. All the fragment peaks are offset by $1Da$ at a time, consecutively up until the highest m/z value in the MS/MS spectrum. Once the offset peaks have reached the highest m/z value, they are rotated back to the lowest m/z value, and further rotated until they have reached their original m/z value. Each $1Da$ shift produces a new decoy spectrum. The shifting method is inspired by the calculation of cross-correlation

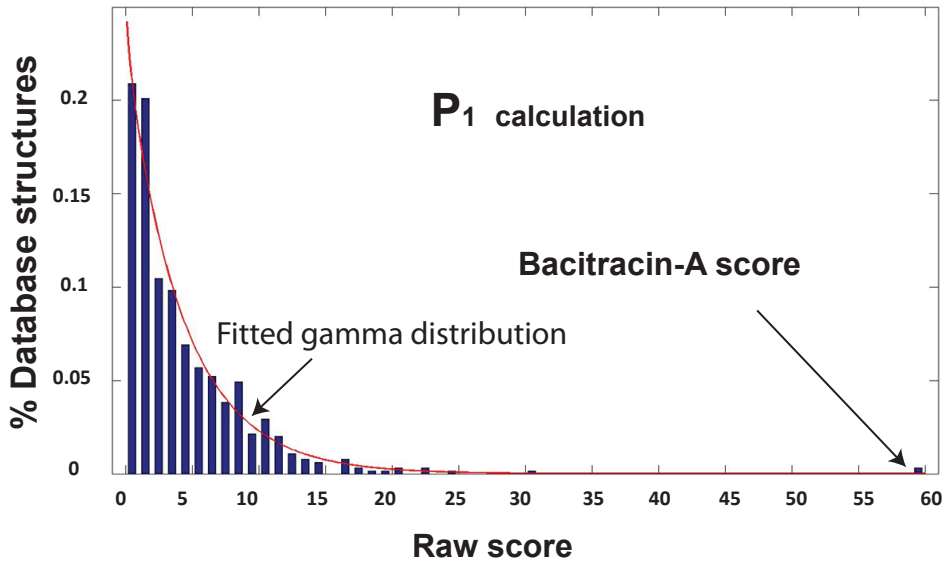


Figure 3.4: Calculation of P_1 score on a bacitracin-A MS/MS spectrum.

score in the SEQUEST algorithm, which was the first computer algorithm for matching ribosomal peptides in a database with the MS/MS spectral data.

The P_2 score is calculated in a similar fashion to that of P_1. However, the gamma distribution of the raw scores is estimated from the raw scores between the decoy spectra and the candidate structure. The p-value is the exceedance frequency at the original MS/MS spectrum's raw score. (See Figure 3.5)

The p-value represents the probability of that a random spectrum being scored higher with the candidate structure than the original MS/MS spectrum. A low p-value indicates a greater probability of the spectrum being correctly matched. A P_2 score is derived from the p-value, whereby, a higher value represents a greater significance of being correctly matched.

$$P_2\ score = -10\,log_{10}(p\text{-}value)$$

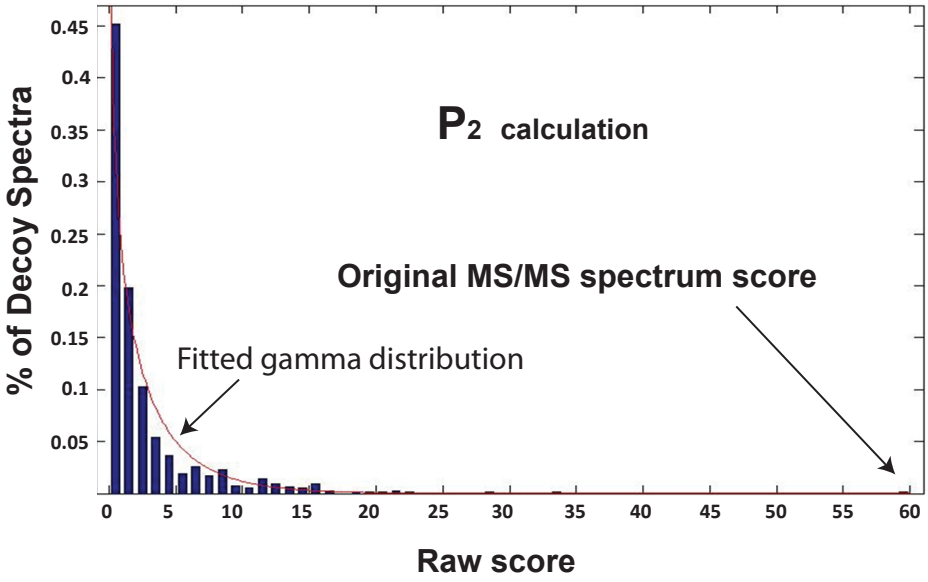


Figure 3.5: Calculation of P_2 score on a bacitracin-A MS/MS spectrum.

3.3.4 P_1, P_2 Thresholds and Result Filtering

iSNAP analyzes each input MS/MS spectrum individually, and the scoring scheme is applied. As mentioned previously, NRPs within the mass range of $[M] \pm 1Da$ are considered as candidates for an input spectrum. If no NRP in the database is found within the $[M] \pm 1Da$ mass range, then no candidate can be established. As such, there would no identification for with that MS/MS spectrum. For an MS/MS spectrum with candidates, a P_1 score and a P_2 score is generated for each candidate.

iSNAP applies thresholds on P_1 and P_2 scores to judge whether a candidate is significant enough to be reported. The threshold values are determined by threshold training experiments, which are explained in Experiment II. A candidate compound is regarded as identified only if it has both P_1 and P_2 scores above the corresponding thresholds and also have the highest P_1 score among all the candidates for that spectrum.

To ensure the input MS/MS is well matched to the compound, additional filtering is applied to the candidates above thresholds. A candidates is discarded if its hypothetical spectrum matches the input spectrum with less than 5 peaks. On top of that, a candidate is also discarded if number of the matched peaks are less than 10 while more than 75% of

matched peaks have relative intensity less than 2%.

In the output, iSNAP assembles identification report for each MS/MS spectrum; outlining the scan number, retention time (if applicable), precursor m/z, charge state, precursor's mass, as well as information about all the candidates for the spectrum, including the candidate's name, SMILES code, monoisotopic mass, the number of fragments matched, raw score, P_1 score and P_2 score. Figure A.2 is the screenshot of a typical identification report.

3.4 Experiments and Results

3.4.1 Experiments Overview

In this section, we present five experiments on iSNAP database search. The first two experiments were designed to support the development of iSNAP, and the following three aimed to evaluate iSNAP's performance with progressively challenging settings.

Experiment I checked the scoring scheme's capability in ranking the NRP candidates for a single MS/MS spectrum. In order to make a correct identification, it is essential for the scoring scheme to distinguish the genuine candidate from all structures in the database. This experiment tested iSNAP's scoring scheme with the MS/MS spectra of bacitracin-A. Each bacitracin spectrum was scored with all NRPs in the database. With the distribution of P_1 and P_2 scores over all database NRPs, we examined whether the true structure bacitracin-A could be scored distinguishably higher than other NRPs.

Experiment II aimed to establish appropriate thresholds for P_1 and P_2 scores. The thresholds are used to determine whether a top candidate on the spectrum is significant enough to be reported as an identification. The top candidate is the candidate NRP with highest P_1 score, which is considered the best match for the spectrum. However, it does not necessarily imply the match is true, because the true structure may not exist in the NRP database. If the top candidate's P_1 or P_2 score is low, it is more likely the top candidate turns out to be a false match. It is important to set common criteria in order to interpret the P_1 and P_2 scores across all the MS/MS spectra.

Furthermore, we demonstrate the effectiveness of iSNAP database search with progressively challenging experiments. In Experiment III, iSNAP was evaluated to identify spiked NRPs within complex mixtures. A mixture of six standard NRPs were spiked into 11 different fermentation media. LC-MS/MS data of each medium was acquired using data dependent acquisition. iSNAP was used to detect the spiked NRPs from the 11 LC-MS/MS dataset. As the fermentation media contained varying amounts of peptides and proteins

but no NRP structures, it provided an ideal matrix in which to test the false positive and false discovery rates.

Experiment IV and V were set up to demonstrate that iSNAP could detect naturally produced NRPs. The experiments were designed to emulate how the software would be used by lab researchers. In Experiment IV, the bacteria *kutzneria sp.744*[39] was cultured in regular medium to produce kutzneride. Kutzneride is a highly modified nonproteongenic nonribosomal peptide with complex structure. The LC-MS/MS of the extract was analyzed by iSNAP to show the compound can be correctly identified. Moreover, cultured in brominated medium, *kutzneria sp.744* was anticipated to produce di-bromo-kutzneride, which was not in the NRP database, but the structure could be predicted. By including the brominated structure, di-bromo-kutzneride, into the NRP database, iSNAP identified it from the MS/MS spectra and helped confirm its existence. In Experiment V, *Bacillus sp.*[16] was cultured to produce tyrocidines, which is a series of bioactive cyclic NRPs with closely similar structures. iSNAP was used to interrogate the LC-MS/MS of the microbial culture, and distinguishably identified a series of compounds in tyrocidine family.

The mass spectral data used in this research were generated by research collaborators in Nathan Magarvey lab in McMaster University, using a Bruker amaZon-X ion-trap instrument with electro-spray ionization source. The mass spectrometer was coupled to a Dionex Ultimate 3000 HPLC system to perform LC-MS/MS analysis.

3.4.2 Experiment I - Validation of P_1 and P_2 Scoring

The experiment aimed to validate that the scoring scheme is capable of distinguishing the true candidate from 1107 database NRPs for an input MS/MS spectrum. To evaluate the scoring scheme, an MS/MS spectrum of doubly protonated (711.82 m/z) bacitracin-A was selected as the input. The spectrum was searched with the database, and P_1 And P_2 scores were calculated for each database NRP. As the true candidate bacitracin-A is in the database, it was expected that bactracin-A would have a distinctively higher P_1 and P_2 score than those of other database NRPs.

We sorted the 1107 database NRPs by their P_1 scores. Table 3.1 shows the top five NRPs for the bacitracin-A spectrum. It is noticed that the true candidate bacitracin-A was scored distinguishably higher in both P_1 and P_2, which are 57.6, 72.3, respectively. The second ranked NRP, bacitracin-F, is also in the bacitracin family with a similar structure to bacitracin-A, as shown in Figure 3.6. It was scored with P_1=29.8, P_2=31.7, which was higher than the others but substantially lower the scores of the true candidate. The result

Rank	Database NRP	Mass	Raw score	P_1 score	P_2 score
1	Bacitracin_A	1421.748	59.09543	57.5842	72.31199
2	Bacitracin_F	1418.701	30.48299	29.84443	31.7187
3	Tyrocidine-D	1370.692	24.79772	24.32377	18.72659
4	Tridecaptin_B_gamma	1459.803	22.1822	21.78215	18.56334
5	Petriellin_A	1430.878	22.10258	21.70476	31.00149

Table 3.1: P_1, P_2 scores of the top five database NRPs for a bacitracin-A MS/MS spectrum. The bacitracin-A MS/MS spectrum is searched with the database of 1107 NRPs. The true candidate bacitracin-A is scored at the top, distinguishably higher in both P_1 and P_2 score. The second ranked NRP, bacitracin-F, has the similar structure to bacitracin-A. It is scored higher than the others but substantially lower than bacitracin-A.

illustrates that the scoring scheme is usable in distinguishing the true candidate from the database, even from NRPs with similar structures.

To further demonstrate this point, a working standard of bacitracin-A was infused directly into the mass spectrometer for approximately 1min. This MS/MS experiment generated 56 bacitracin-A scans. The test was repeated on each of those scans. For each scan, the P_1 and P_2 scores of the true candidate, bacitracin-A, was recorded, represented by a blue point in Figure 3.7. The scores of other database NRPs are represented by red points. As such, the results on 56 bacitracin-A scans are overlaid in Figure 3.7, which shows bacitracin-A is consistently scored higher, compared to other database structures.

3.4.3 Experiment II - Determination of P_1 and P_2 Score Thresholds

This experiment aimed to determine the values of P_1 and P_2 thresholds. iSNAP uses the thresholds as criteria by which it decides whether to report an MS/MS spectrum's top candidate as an identification. When an MS/MS spectrum truly corresponds to a database NRP, the P_1 and P_2 scores should be high, as shown in the previous experiment. However, an MS/MS spectrum of other substance outside the database can still have candidates in the $[M] \pm 1Da$ mass window by chance. Normally, theoretical spectra of these false

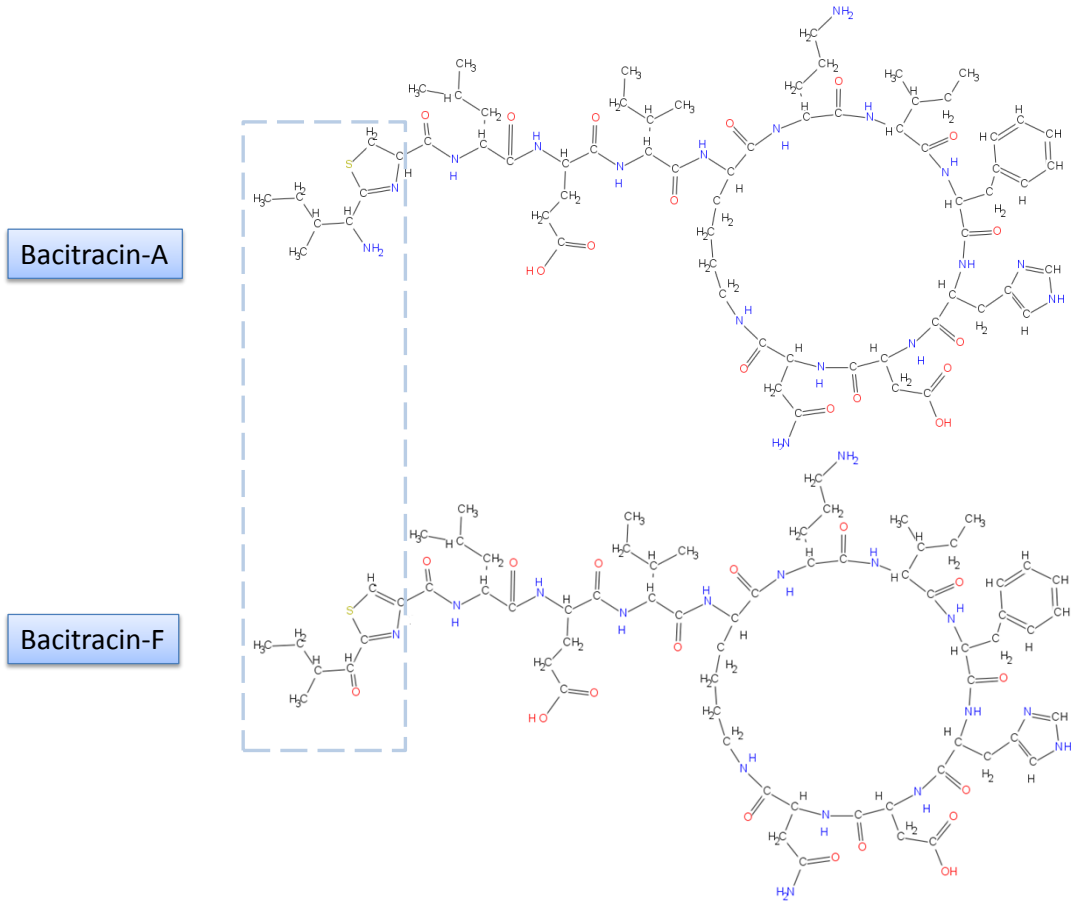


Figure 3.6: Structural comparison of bacitracin-A and bacitracin-F. The two structures are almost identical except the difference in the building block at branch terminal.

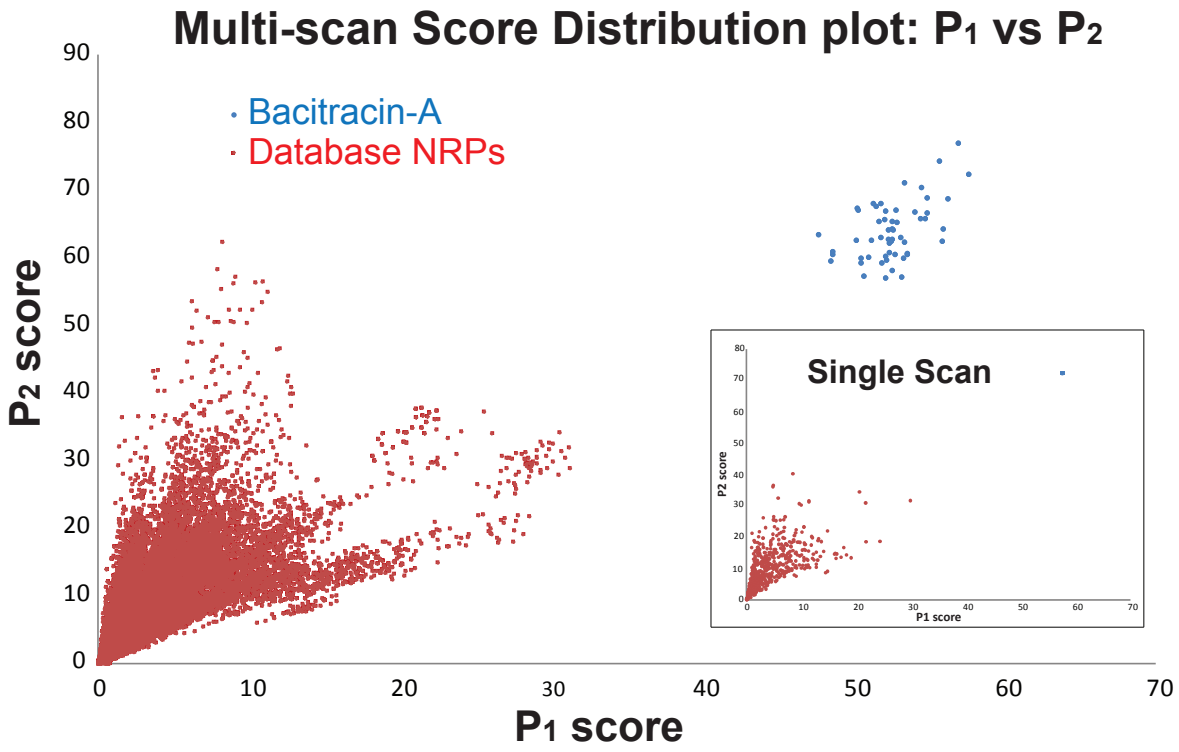


Figure 3.7: Score distribution of the 1107 database NRPs on 56 bacitracin-A MS/MS spectra. The P_1 and P_2 scores of the true candidate, bacitracin-A, are represented by a blue points. The scores of other database NRPs are represented by red points. The true candidate bacitracin-A is consistently scored higher in both P_1 and P_2.

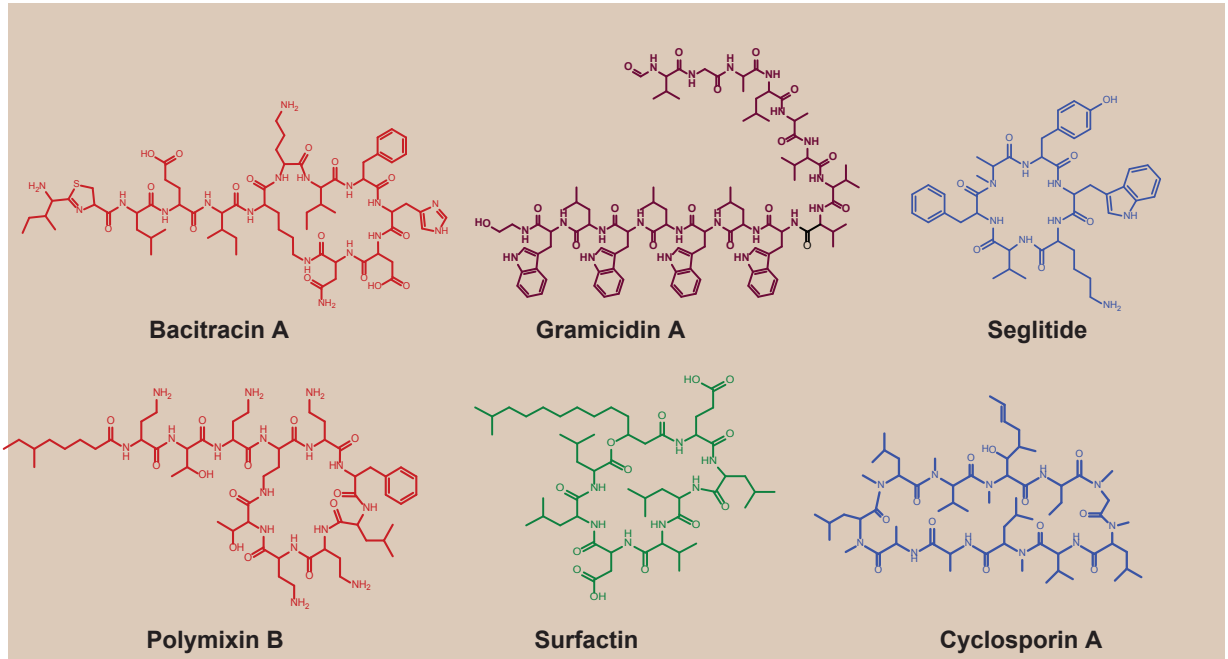
candidates cannot match well with the input spectrum. Therefore, even the top candidate should have low scores.

Higher P_1 and P_2 scores generally imply a greater probability for the top candidate to be true. Ideally, appropriate thresholds is supposed to optimally separate the true and false top candidates. To establish appropriate thresholds, it is necessary to know the P_1 and P_2 distributions of the true and the false.

The scores of true top candidates were generated using the MS/MS spectra of six different NRPs. The six NRPs are in the database, including bacitracin-A, cyclosporin-A, gramicidin-A, polymyxin-B, surfactin and seglitide. They represent typical NRP structures (See Table 3.2). For each NRP, a set of MS/MS spectra were generated by infusing the NRP working standard to the mass spectrometer. A total of 367 MS/MS spectra were generated from the six compounds. iSNAP searched these spectra with the NRP database. All spectra had the correct NRP scored as the top candidate. The P_1 and P_2 scores of these true top candidates were acquired, and used as positive controls in Figure 3.8. In Table 3.2, the structures of the six NRPs are illustrated and the highest P_1 and P_2 scores for each compound are listed.

The scores of false top candidates were generated with the LC-MS/MS data of 11 common fermentation media. The 11 media are blank controls with no NRPs existed. A total of 12569 MS/MS spectra were acquired by data dependent acquisition. Their precursor mass distributed over a range from 300 to 2700 m/z. Processed with iSNAP database search, 6744 out of 12569 MS/MS spectra had least one candidate from the NRP database. As no NRPs existed within those media, the top candidate must be falsely matched to the spectrum. Their P_1 and P_2 scores were generally low, and are used as the negative controls in Figure 3.8.

The score distribution of the true and false top candidates are shown in a P_1-P_2 scatter plot (Figure 3.8). We empirically estimated the P_1 threshold at 27, corresponding to a p-value at 0.002, and the P_2 threshold at 24, corresponding to a p-value at 0.004. The thresholds balanced the sensitivity and specificity at 91.5%, 99.8%, respectively. It allowed a sufficient amount of true candidates to be reported, while kept a low ratio of false positives. Using the estimated thresholds, 335 of 367 true top candidates could be positively identified, while 24 of 6744 top candidates from media spectra became false positives.



Structure	Candidate	Precursor m/z	Charge state	Precursor mass	Candidate mass	P1 score *	P2 score
Linear	Gramicidin A	1882.5	1	1881.5	1881.1	34.6	40.7
Linear-Cyclic	Bacitracin A	711.9	2	1421.7	1421.7	57.6	72.3
Linear-Cyclic	Polymixin B	602.0	2	1202.0	1202.7	34.6	35.0
Cyclic	Surfactin	1037.2	1	1036.2	1035.7	28.5	31.2
Cyclic	Cyclosporin A	1203.4	1	1202.4	1201.8	35.1	41.8
Cyclic	Seglitide	405.5	2	809.0	808.4	57.5	48.2

* Multiple MS/MS scans are analyzed for each compound. The scores of the scan with the highest P1 is presented.

Table 3.2: Structures and P_1, P_2 scores of six representative NRPs.

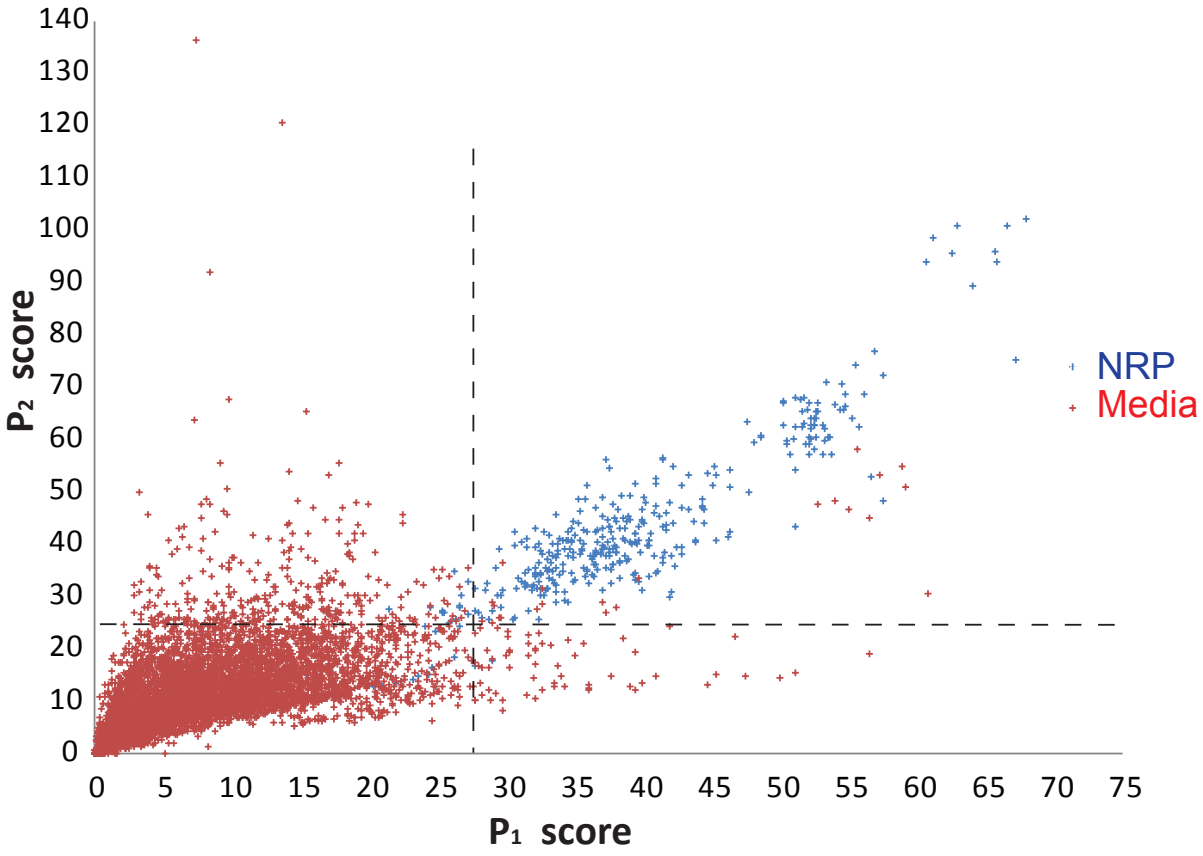


Figure 3.8: P_1-P_2 distribution of true and false top candidates for threshold determination. The scores of true candidates were generated with MS/MS spectra of six standard NRPS. The scores of false candidates were generated with 11 different media panel. Thresholds on P_1 and P_2 are empirically determined at 27 and 24, respectively.

3.4.4 Experiment III - Identification of Six Spiked NRPs within Complex Fermentation Media

In this experiment, the determined P_1 and P_2 thresholds were examined in the analysis of LC-MS/MS datasets from complex mixtures, in order to check their ability in separating the true and false top candidates. The thresholds are used as criteria for deciding whether a top candidate should be reported as an identification. It is important to have an estimation about false discovery rate [40] and false positive rate when the determined P_1, P_2 thresholds are applied, in the analysis of complex samples.

The thresholds were evaluated with spike-in studies of 11 fermentation media, each spiked with the mixture of six NRPs. For each spiked NRP, the final concentration was set to $50\mu g/mL$. The panel of 11 spiked fermentation media was analyzed with LC-MS/MS, where tandem spectra were generated using data dependent MS/MS acquisition.

Automated MS/MS acquisition generated a total number of 6793 MS/MS spectra from the 11 fermentation media. The algorithm analyzes these MS/MS spectra with our scoring scheme. It turned out 4198 out the 6793 MS/MS spectra has at least one candidate with a $\pm 1Da$ precursor mass window. For the 4198 scans with candidates, the algorithm matches the candidate with highest P_1 score to the spectrum. We call such a match as a PSM (peptide-spectrum match). A PSM is true if the top scored candidate is the genuine structure for the spectrum, otherwise the PSM is false. In this sense, the 4198 scans with candidates yielded 4198 PSMs, each with a P_1 score and a P_2 score. We ignored the rest of 2595 scans with no candidate, as they must correspond to compounds outside the NRP database. The database search algorithm can not make identifications if the compound is not in the search scope.

To estimate the false discovery rate and the false positive rate, we let the *null hypothesis* be a false PSM having both P_1 score and P_2 score above the thresholds, which is the case that an MS/MS spectrum is wrongly matched to a database NRP but the algorithm considers it as significant and report it as an identification. We perform multiple hypothesis tests on PSMs by running this experiment.

In Figure 3.9, the 4198 PSMs are plotted on P_1 and P_2 scores. Further analysis showed that 1455 of the 4198 PSMs were matches to one of the six spiked NRPs. Those PSMs are marked accordingly. PSMs not matched to any one of the spiked NRPs are marked with red dots.

To validate whether the 1455 PSMs were true, we inspected how well these MS/MS spectra were matched. For each spiked NRP, its MS/MS spectra at the same charge state are generally consistent with each other across the 11 datasets, as the experimental settings

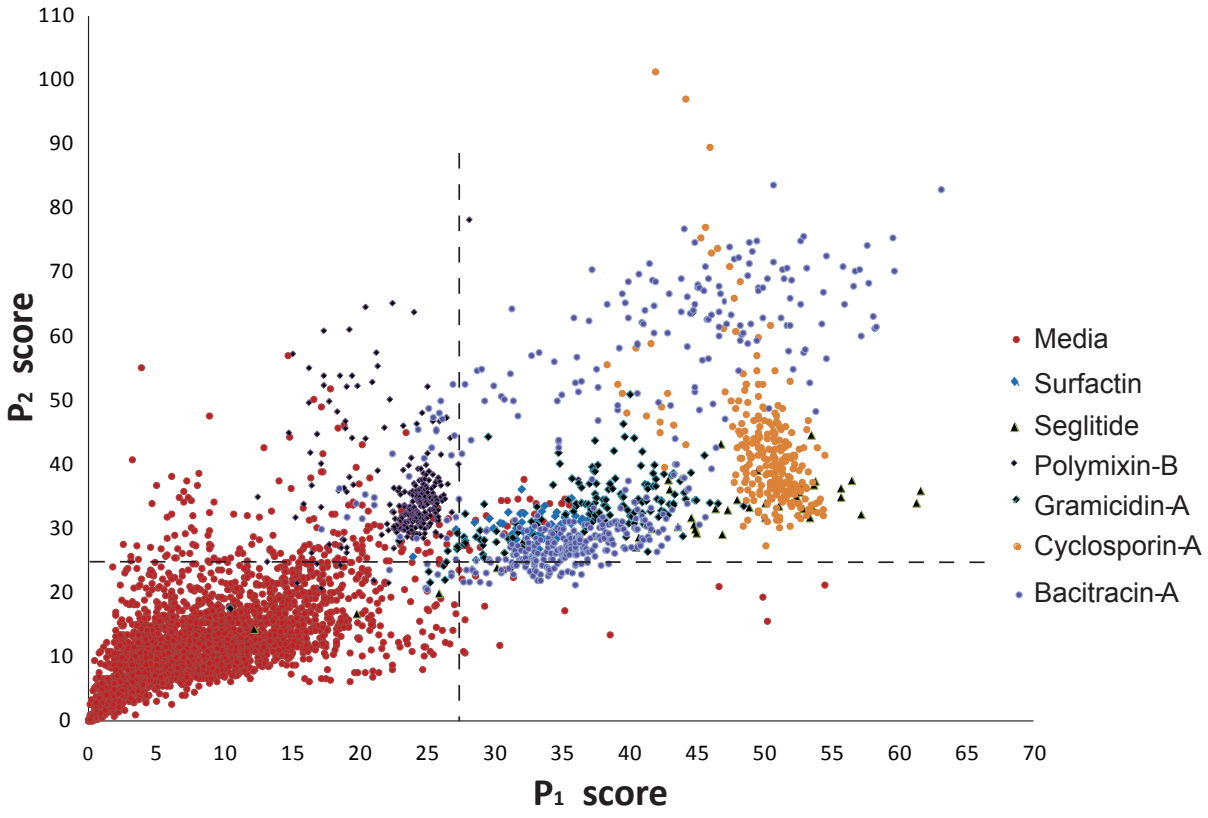


Figure 3.9: P_1-P_2 distribution of 4198 PSMs in the analysis of 11 NRP-spiked fermentation media. Each point in the figure indicates a PSM (peptide-spectrum match), which is the match of an MS/MS spectrum and its top scored candidate. 1455 of the 4198 PSMs are true PSMs to one of the six spiked NRPs. The other 2743 PSMs are false, marked as red dots. The majority of PSMs to the six spiked NRPs scored above the P_1 P_2 thresholds except matches to polymyxin-B.

were the same. A match was considered to be true if the MS/MS spectrum conformed with common fragmentation pattern of the matched NRP. The 1455 PSMs were manually examined, and all of them were deemed to be true.

The rest of 2743 PSMs not matched to spiked NRPs were determined to be false. Theoretically, these PSMs could be categorized to two cases. In the first case, the MS/MS spectrum was indeed a genuine scan of a spike NRP, but the algorithm wrongly ranked another NRP to be the highest scored candidate. In the second case, the spectra could correspond to other substances in the medium. However, it has been known that there are no NRPs in the medium, therefore, any candidate from the NRP database matched to the spectrum must be false. In conclusion, the rest of 2743 PSMs were all false.

At this moment, the distinguishing power of P_1 and P_2 thresholds could be evaluated. The P_1 and P_2 thresholds had been previously established as 27 and 24, respectively, and was applied to the 4198 PSMs in Figure 3.9. For clarity, we define an *identification* as a PSM with P_1 and P_2 above thresholds.

We checked the false discovery rate and false positive rate with the thresholds. In this case, false discovery rate is the number of false *identifications* over all *identifications*. These were 23 false PSMs above both the thresholds, which were the false *identifications*. Meanwhile, 992 of the 1455 true PSMs were *identified*. This yielded a false discovery rate of $\frac{23}{23+992} = 2.27\%$. False positive rate is the number of false *identifications* over all false PSMs, which was calculated as $\frac{23}{2743} = 0.84\%$.

Table 3.3 shows the false discovery rate and false positive rate calculated separately on each of the 11 datasets.

There is another indicator, true positive rate, which is calculated as the number of true *identifications* over all true PSMs. A higher true positive rate suggest less true PSMs will be missed in the identification report. This is thought less important with the reason that dereplication can be made as long as one of true PSMs has the NRP scored above the thresholds.

As for the identification of the six spiked NRPs, the majority of PSMs to the six spiked NRPs could be identified from the 11 fermentation media except polymixin-B. In TSB and LB media only 4 out of the six spiked NRPs could be identified, while 5 out of the six were identified in YPD, YMG, GYM, Pharmamedia, Grasseed, Fishmeal and in CY. In the case of polymixin-B, earlier MS/MS infusion data generated P_1 and P_2 scores above the thresholds. However, it was not dereplicated in most media under automated data dependent acquisition. We speculate this is due to matrix effects in which the coextracted substance might have altered the signal response, which affected the spectral quality of polymyxin-B scans.

Fermentation Media		# MS/MS Scans	# PSMs	# true PSMs	# true PSMs above thresholds	# false PSMs	# false PSMs above thresholds	False Discovery Rate	False Positive Rate
1	YPD	774	467	75	48	392	1	2.04%	0.26%
2	YMG	663	403	87	67	316	0	0.00%	0.00%
3	GYM	440	259	130	94	129	3	3.09%	2.33%
4	TSB	699	402	67	61	335	3	4.69%	0.90%
5	LB	736	443	88	61	355	4	6.15%	1.13%
6	Nutrient	709	493	200	86	293	3	3.37%	1.02%
7	Pharmamedia	423	231	140	106	91	2	1.85%	2.20%
8	Grasseed / Veg.Protein	603	466	145	118	321	0	0.00%	0.00%
9	Fishmeal	496	296	205	152	91	2	1.30%	2.20%
10	R2A	518	312	192	126	120	1	0.79%	0.83%
11	CY	732	426	126	73	300	4	5.19%	1.33%
Summary		**6793**	**4198**	**1455**	**992**	**2743**	**23**	**2.27%**	**0.84%**

Table 3.3: Detailed result on 11 spiked fermentation media with false positive rate and false discovery rate calculated.

The experiment demonstrated that the determined thresholds are capable of separating the true and false top candidates, with the highest false positive rate less than 3% and false discovery rate less than 7% across 11 media. This indicates iSNAP algorithm has the capable of dereplicating a mixture of NRP compounds from complex matrices. The processing speed for the 11 LC-MS/MS datasets is approximately 15 MS/MS scans per second. This makes iSNAP a practical tool for automated high throughput dereplication of NRP compound in complex mixtures.

3.4.5 Experiment IV - Identification of Kutzneride and Di-bromo-kutzneride

In this experiment, iSNAP was challenged with naturally produced NRPs within crude extract. The experiment emulated a lab research in which the bacteria *kutzneria sp.744* [39] was cultured to produce an antibiotic NRP kutzneride. Kutzneride is a cyclic nonribosomal peptide bearing a high degree of non-proteinogenic amino acids. iSNAP was utilized as a software tool to analyze the LC-MS/MS acquisitions from the fermentation extract and to confirm the existence of kutzneride. It is difficult to the detect naturally produced kutzneride as the compound is produced in low abundance[41]. Moreover, by brominating the regular medium, we aimed to produce a novel analog to kutzneride with *kutzneria sp.744*. With the assumption that the bacteria can incorporate brominated nutrients in the

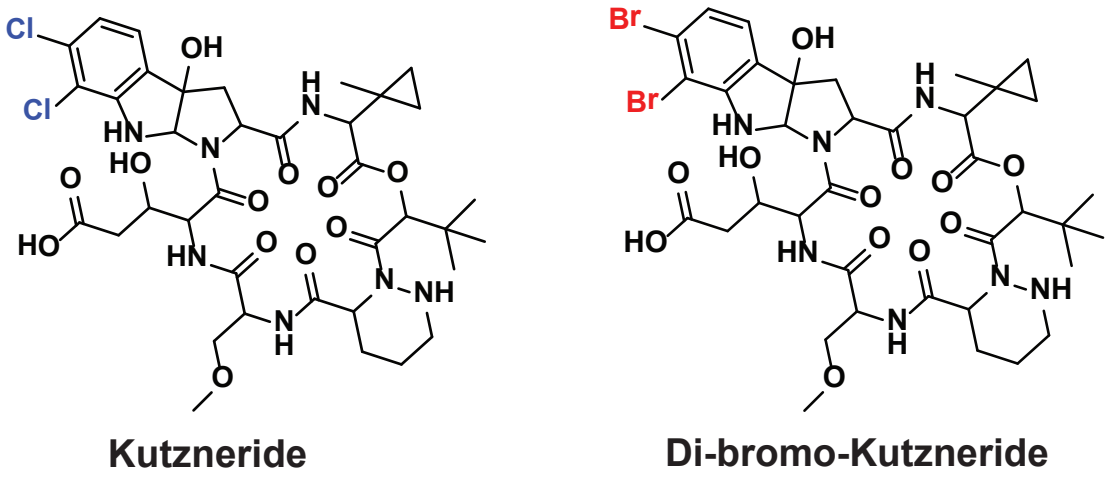


Figure 3.10: Molecular structures of kutzneride and di-bromo-kutzneride.

biosynthesis of kutzneride, it is expected to produce di-bromo-kutzneride. The compound is a close analog to kutzneride by replacing the two chloride groups with bromine groups. Figure 3.10 shows the molecular structures of kutzneride and di-bromo-kutzneride. iSNAP is used to verify the assumption by checking the existence of di-bromo-kutzneride with MS/MS.

The bacteria was cultured by our research collaborators in Nathan Magarvey lab at McMaster University. In the experiment, *kutzneria sp.744* was first grown on ISP-2 agar plates for 7 days prior to a 14-day culturing in Melin Norkrans medium. Culture supernatants were extracted with a HP-20 resins, and subjected to chemical solvent partitioning. The 100% organic fractions where collected and subjected to LC-MS/MS analysis.

LC-MS/MS data files were processed by iSNAP. Kutzneride is in the database of 1107 NRPs. It was identified by the software from the crude organic extract with 4 main fragment peaks (837.29, 836.29, 743.23 and 609.20m/z) being matched with the theoretical spectrum, resulting in highest P_1 and P_2 scores of 31.3 and 33.4 respectively. This initial test was done to reveal that iSNAP could dereplicate kutzneride, but also importantly did not result in any false positives identified in the extract.

A second test was conducted to determine if the iSNAP algorithm could be used to confirm the production of di-bromo-kutzneride in brominated medium. To obtain the novel analog of kutzneride, Melin Norkrans medium is modified using a series of bromide salts. As

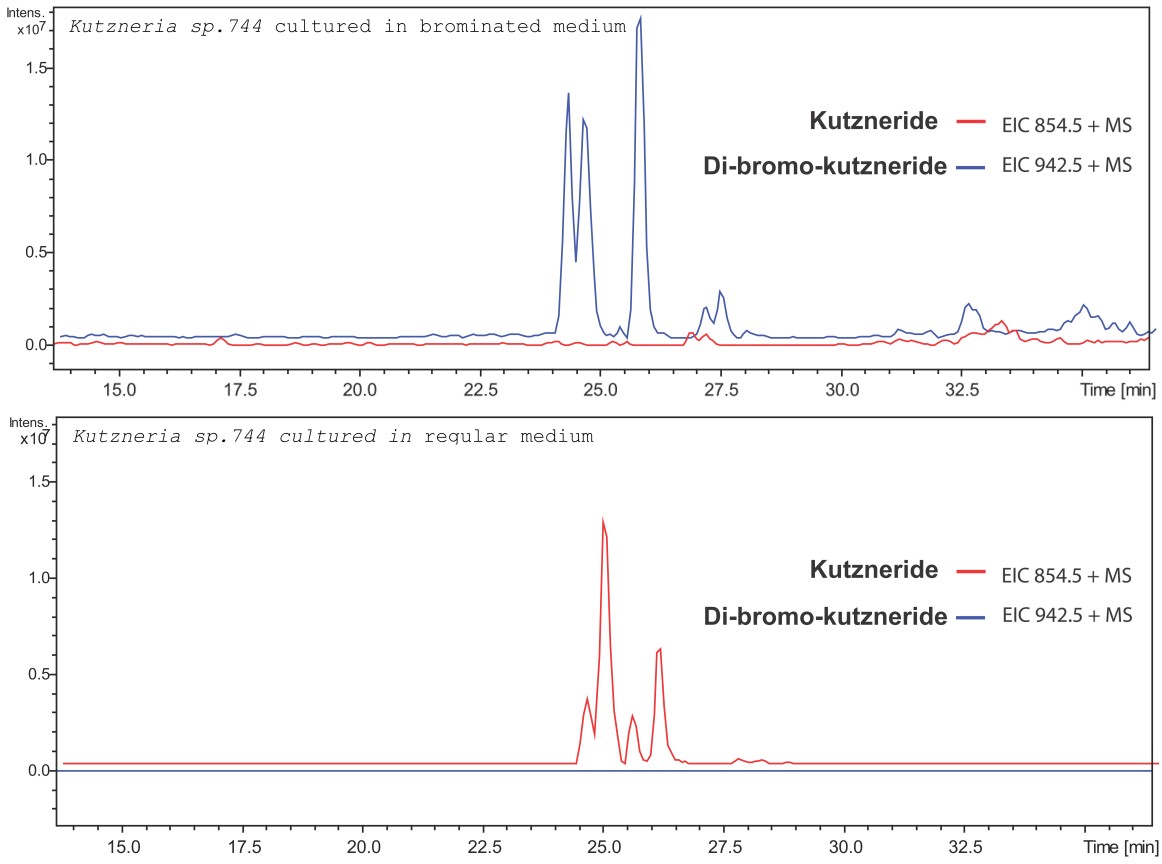


Figure 3.11: Ion counts in the mass window of kutzneride and di-bromo-kutzneride over LC retention time. The bottom chart is acquired with LC-MS on the fermentation using regular medium, implied kutzneride was possibly produced and indicated di-bromo-kutzneride can not be produced in the regular medium. The top chart is acquired with LC-MS on the fermentation using brominated medium. With a substantial increase of ion count in the mass window of di-bromo-kutzneride, it was highly possible the brominated compound was produced. iSNAP was used to structurally confirm the existence of kutzneride and di-bromo-kutzneride, respectively, with MS/MS spectra.

the mass of di-bromo-kutzneride was easily calculated, we checked the mass on the LC-MS chromatogram of the resulting extract. The chromatogram implied the possible presence of di-bromo-kutzneride with a molecular weight of 942.1 $[M + H]^+$ and a decrease in the production of kutzneride (See Figure 3.11 Top). A series of MS/MS scans were acquired using the calculated mass as the manually selected acquisition window. iSNAP analyzed the MS/MS dataset with the database of 1107 NRPs and reported no identification, as di-bromo-kutzneride was not in the NRP database, we constructed the molecular structure in ChemDraw, and exported the SMILES code. After adding the SMILES code of di-bromo-kutzneride to the database, iSNAP was used to re-analyze the MS/MS data. A total of 4 high intensity fragment peaks where matched form the MS/MS spectra (942.18, 925.18, 924.18 and 830.12 m/z). The highest P_1 and P_2 scores on the MS/MS spectra were 75.9 and 29.3 respectively. The matched peaks represented the fragments of di-bromo-kutzneride, which confirmed that a correct identification was made.

3.4.6 Experiment V - Identification of a Series of NRPs in Tyrocidine Family

In this experiment, iSNAP is tested to identify a series of naturally produced NRPs with similar structures. It is mention previously that NRPs in a same family can be produced simultaneously in a biosynthesis pathway, resulting structurally similar NRPs in the same microbial fermentation. As an example, tyrocidine family contains more than 28 antibacterial cyclic NRPs[16], and they only different at a few amino acid residues (Figure 1.4). Out of the 28 known tyrocindines, only five of them, tyrocidine A, B, C, D and E, have been conclusively characterized, and have the SMILES code in our NRP database. This phenomenon not only gives a good reason to develop an NRP analog search algorithm, but also provides a decent challenge for iSNAP database search. In order to be a practical tool for NRP dereplication, iSNAP should be able to simultaneously and distinctively identify those similar NRPs in the same LC-MS/MS dataset.

In Nathan Magarvey lab at McMaster University, bacitracin strain *Bacillus sp.*[16] was cultured to produce tyrocidines. In the course of screening the microbial culture for the antibacterial compounds, an assay was used to detect antibiotic agents from crude fermentation extracts. The crude extract was separated by HPLC (Figure 3.12), fragmented a 96 well plate for antibacterial testing, and a bioluminescent strain of staphylococcus was used as a bioactivity indicator. Figure 3.13 shows the screening result of the 96 well plates. Bioactivity was revealed in well C11, D1-6, D8, E1 and F2. The high number of actives from this crude extract made it a suitable material to test iSNAP database search for NRP dereplication.

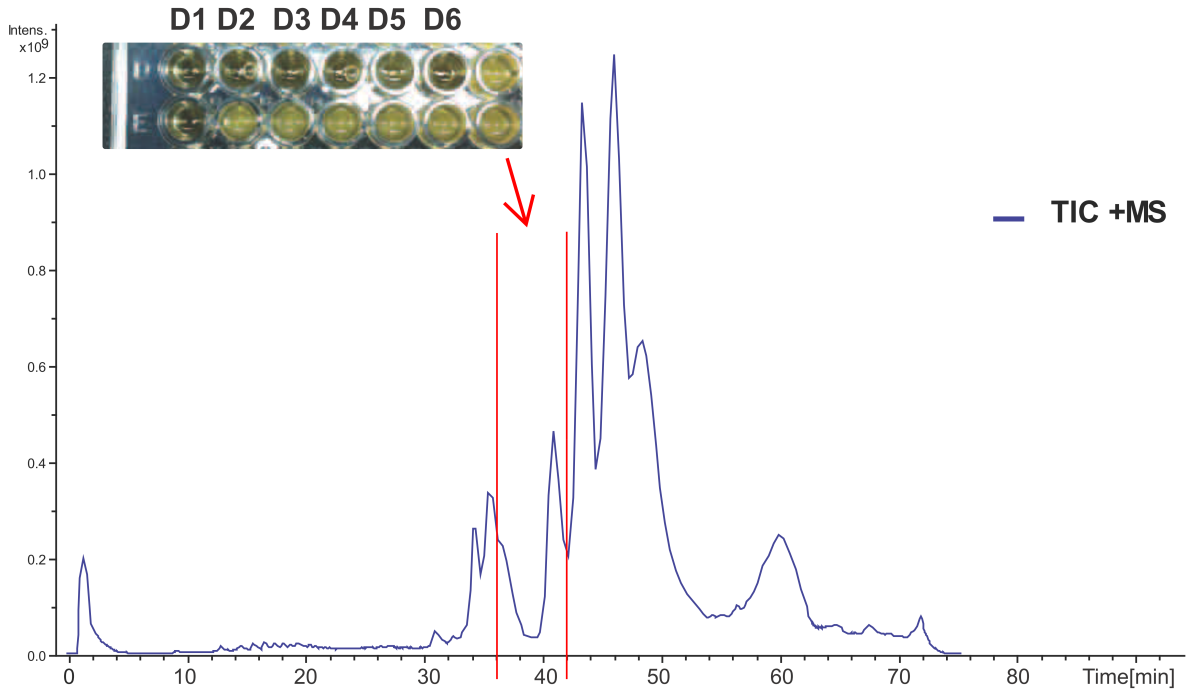


Figure 3.12: Liquid chromatogram of the *Bacillus sp.* extract. The bioactive fractions D1-D6 corresponds to the retention time of 36 - 42min. The LC chart shows the fermentation is a complex mixture with various compounds.

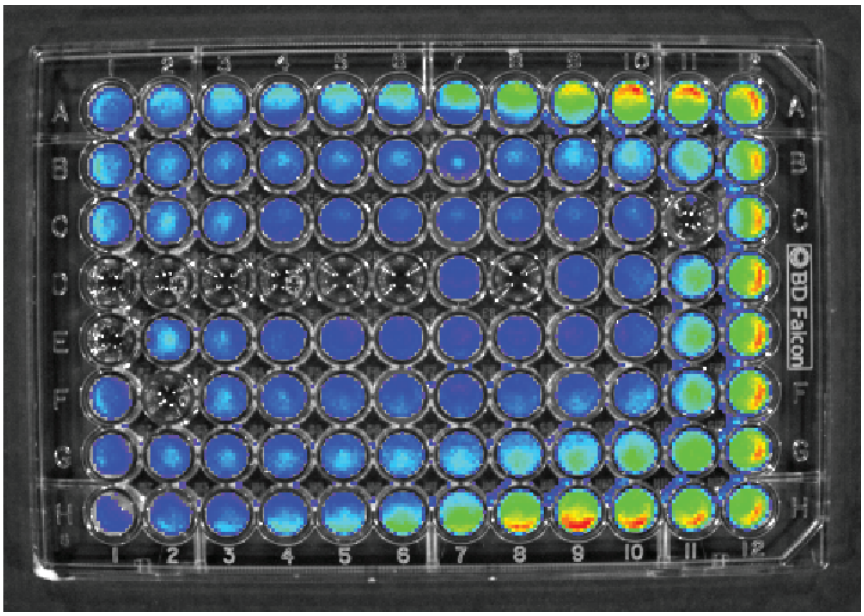


Figure 3.13: Bioactivity screening of LC fractions of *Bacillus sp.* extract. A bioluminescent strain of staphylococcus was used as the bioactivity indicator. Grey wells (C11, D1-6, D8, E1 and F2) indicates that the fractions are antibacterial and have killed the staphylococcus.

Fraction # (96-well)	Rt (min)	ID	Candidate Mass	P1 * score	P2 score
D1,E1	40.5	Tyrocidine A	1269.7	84.7	44.3
D2-D5	38.6	Tyrocidine B	1308.7	86.8	61.6
D2-D6	37.0	Tyrocidine C	1347.7	85.1	44.3
D2-D5	39.0	Tyrocidine D	1370.7	69.5	41.5
D1,E1	40.8	Tyrocidine E	1253.7	72.8	55.0

* Highest P1 result reported

Table 3.4: Identification report of LC-MS/MS of the *Bacillus sp.* extract.

iSNAP was used to interrogate the LC-MS/MS of the microbial culture with the hope to find the tyrocidine A, B, C and D, which are normally the most abundant in the tyrocidine mixtures. 2455 MS/MS spectra were automatically acquired with data dependent acquisition from LC-MS, and were analyzed by iSNAP with the NRP database. As the result, iSNAP not only identified tyrocidine A, B, C, D but also a lesser compound tyrocidine E. Table 3.4 shows the identification report for the five compounds. All of them had P_1 and P_2 scores that were significantly greater than the empirical threshold values of P_1=27, P_2=24.

The LC-MS was reviewed to further validate the identifications of tyrocidine A, B, C, D and E. The LC-MS chromatogram was filtered by the calculated precursor mass from these structures, showing the ion counts of the five tyrocidines (Figure 3.14). The five compounds eluted from 36-42min, with slight differences in retention time, which suggested the eluted compounds probably had similar structures and were in the same family. The observation was consistent with the assumption that tyrocidine A, B, C, D and E existed in the sample.

In this iSNAP analysis, 45 of 2455 MS/MS scans had the top candidate scored over the thresholds, of which 41 belong to the tyrocidine family. The remaining 4 hits were considered false identifications, but all have low P_1 and P_2 scores. The 4 false hits ranked 41-43, 45 after sorting the 45 identifications by P_1 score, and all 40 identifications before it are tyrocidines. In addition, it was noticeable that tyrocidine E was positively identified from the NRP database and dereplicated with high confidence (P_1=72.8 and P_2=55.0), even though its relative abundance was 2-3% to the most abundant tyrocidine B, showing that low abundant NRP could still be dereplicated with iSNAP.

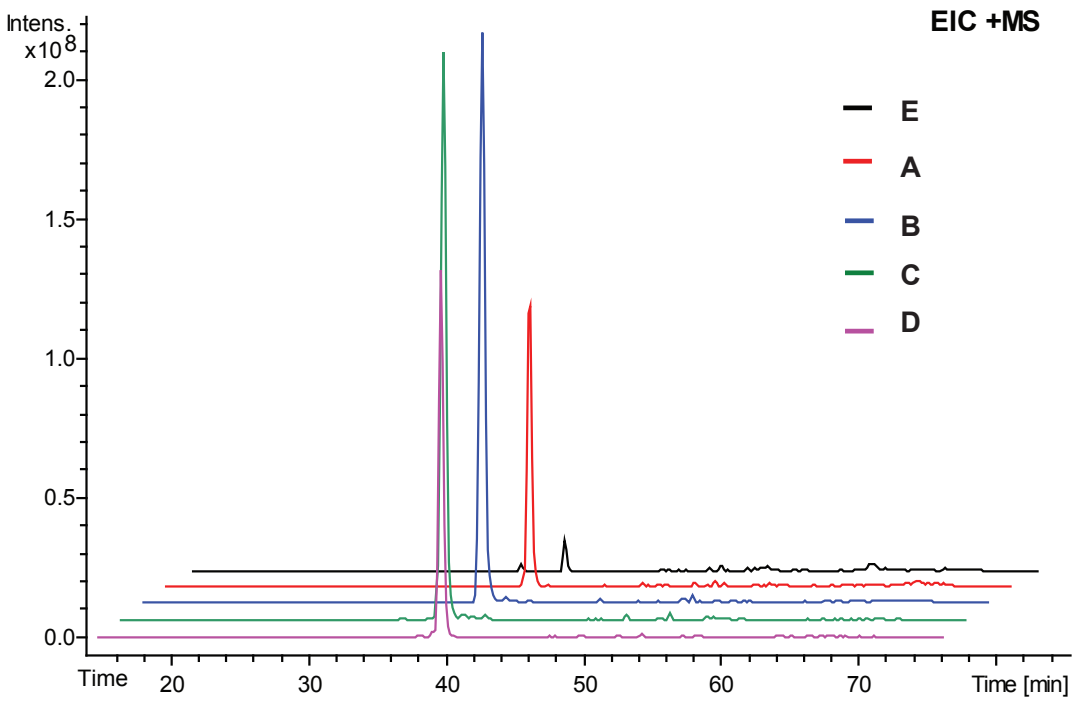


Figure 3.14: Ion counts in the mass window of tyrocidine A, B, C, D, E over LC retention time. The five compounds eluted from 36min-42min, with slight differences in retention time, which suggested they probably had similar structures and were in the same family. The observation was consistent with the assumption that tyrocidine A, B, C, D and E existed in the sample.

Chapter 4

iSNAP for Semi-automated Nonribosomal Peptide Analog Search

4.1 Overview

In the previous chapter, we introduced iSNAP as an informatic tool that identifies nonribosomal peptides from tandem mass spectra. It is tested to be an effective tool and can facilitate NRP dereplication in lab research. However, the database search algorithm has its own limitation that only the NRPs in database can be identified. It is common that NRPs with similar structures co-exist with the identified NRPs. We refer these similar NRPs, which are not in the database, as analogs. In some cases, those analogs need to be dereplicated as well, because they may not be the interesting subjects in a research. On the other hand, if a research aims to find novel analogs in an NRP family, it can be exciting not only to detect but also acquire some structural information of analogs with the LC-MS/MS of crude extract. Nonetheless, in both cases, a search program for identifying analogs is useful.

To meet the needs, we expanded the database algorithm to be capable of identifying NRP analogs. iSNAP analog search is designed as a procedure which follows iSNAP database search. It utilized the precisely identified NRPs in database search as "seeds" to narrow down the search scope. MS/MS spectra in the input are re-analyzed to identify analogs that are different at one building block to a seed NRP. Identified analogs are reported with P_1, P_2 scores, as well as the database NRP from which the analog is derived, along with the location of the different building block.

It is known that most of successful database search algorithms for traditional peptides have been extended to modified peptides [42] [43]. They let the user specify a small list of possible modifications and also the amino acid residues where each modification can be applied. A different approach is used for iSNAP analog search. We directly consider the mass difference, between the scan's precursor and a "seed", as the mass of a possible modification which can convert the "seed" to an NRP analog in the sample. We reason that it is hard to have a selected list of anticipated modifications for NRPs, due to the huge number of modifications that can occur. Our approach is more usable for NRP analog identification, but it is limited to find analogs with only one different residue.

iSNAP analog search can be used iteratively. If some analogs are found using analog search, these identified analogs can be used as "seeds" in the next round of analog search. Therefore, it is possible to find analogs which are different at more than one building blocks.

4.2 Method

The workflow of analog search is illustrated in Figure 4.1. To begin with, the analog search algorithm obtains the list of identified NRPs from database search. As mentioned in the previous chapter, those identified NRPs are the top candidates on their corresponding MS/MS spectrum with P_1 and P_2 scores above the thresholds. The software allows users to filter the list so that only the confidently identified NRPs are used as seeds. We denote the seeds as S_i, $i = 1 \dots n$. For each seed S_i, the algorithm annotates amide bonds in the structure, so that the monomer building blocks between two adjacent amide bonds are detected. We denote these building blocks of S_i as R_{ij}, $j = 1 \dots m$. These building blocks are the potential sites that can be altered to generate analog candidates.

In the process of analog search, the input MS/MS spectra are re-analyzed. For each MS/MS spectrum, the precursor mass m is compared with the mass of each seed NRP, $m(S_i)$, $i = 1 \dots n$. When the mass difference $|m - m(S_i)|$ is smaller than a user-specified Δ, which is $150Da$ by default, iSNAP considers that the MS/MS spectrum could potentially represent an analog to the seed NRP.

Analog candidates for the MS/MS spectrum are generated using S_i with $|m - m(S_i)| < \Delta$. For each seed S_i, its building block R_{ij} are enumerated. By assuming the mass difference $m - m(S_i)$ is caused by the structural change in R_{ij}, iSNAP adds the mass difference to R_{ij}, so that the whole structure can then have same mass as the MS precursor. The structure with an altered building block becomes an analog candidate for the MS/MS spectrum.

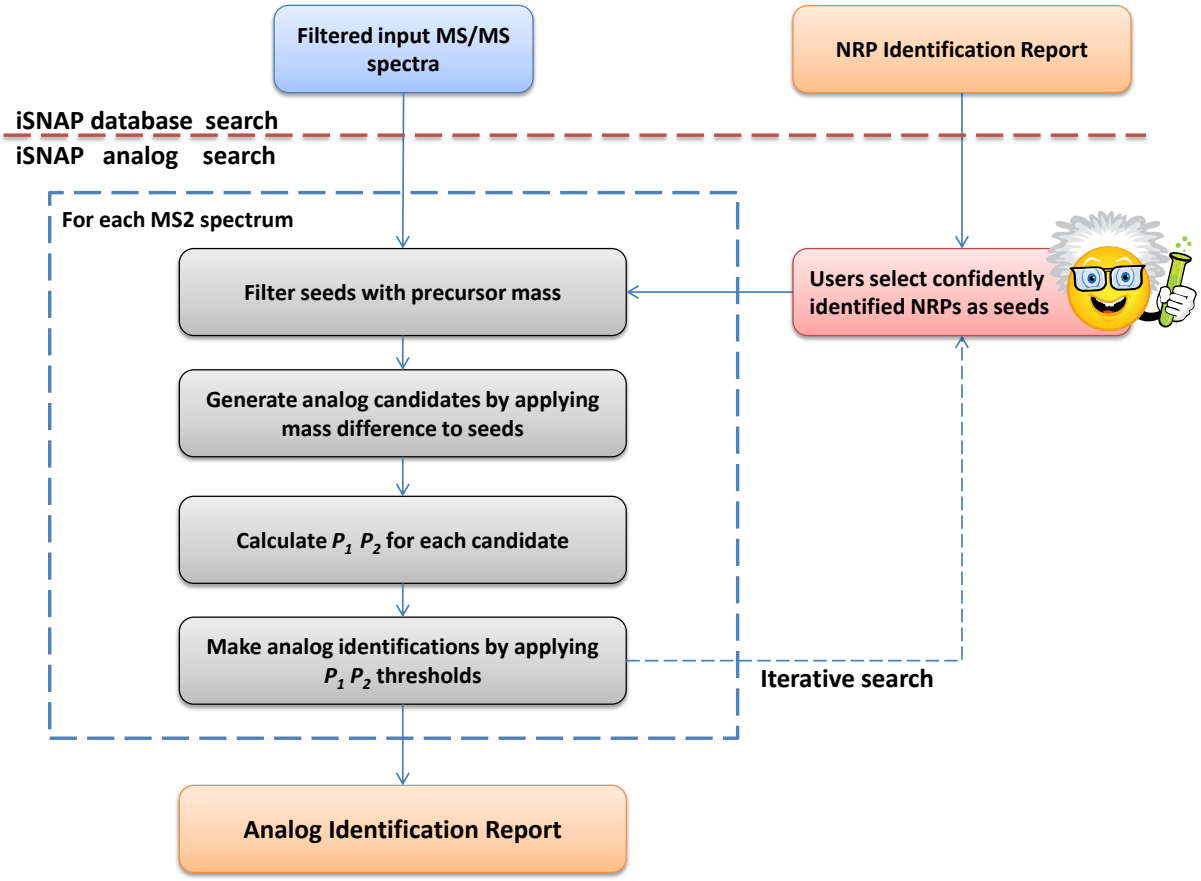


Figure 4.1: Workflow of iSNAP analog search.

The only exception is when $m - m(S_i)$ is negative and $m(R_{ij})$ is too small to account for the mass deduction. With this method, a pool of analog candidates can be generated.

Using the same scoring scheme in database search, P_1 and P_2 scores are calculated for each analog candidate. In the first step, the analog candidate's hypothetical spectrum is generated with *in-silico* fragmentation. iSNAP matches the hypothetical spectrum with the input MS/MS spectrum to calculate a raw score. The raw score is then converted to P_1 score using the raw score distribution which was previously estimated in database search by scoring the input MS/MS spectrum with the hypothetical spectra of database NRPs. To calculate a P_2 score, decoy spectra generated by shifting the input MS/MS spectrum are compared to the hypothetical spectrum of the analog candidate. This generates a raw score distribution, and it is used to convert the original raw score of to a P_2 score.

Finally, an analog candidate is reported as identified if the following criteria are met.

1. The candidate should have the highest P_1 score among all candidates generated for the MS/MS spectum.
2. The candidate should have both P_1 and P_2 scores above the corresponding threshold.
3. The number and quality of matched peaks in MS/MS spectrum should pass the additional filtering described in Section 3.3.4.

In the report, iSNAP outputs the SMILES code of the seed NRP and marks the building block where the mass difference is located.

4.3 Experiments and Results

4.3.1 Experiment VI - Iterative Analog Search for Naturally Produced Tyrocidines

The purpose of this experiment is to evaluate the usefulness of iSNAP analog search in identifying NRP analogs within crude fermentation extract. iSNAP analog search was performed on the LC-MS/MS data previously examined in Experiment V. The LC-MS/MS data was acquired from crude extract of *Bacillus sp.* fermentation and contains various compounds in the tyrocidine family. With database search, all five tyrocidines in the database, tyrocidine A, B, C, D and E were identified. We argue that tyrocines are excellent

testing material for this analog search algorithm, as it is common that compounds in tyrocidine family are only different at one amino acid residue from each other.

Figure 4.2 compares the molecular structures of tyrocidine A, B, C, D and E. It is noticed that,

- Tyrocidine-A ($1269.7Da$) $\rightarrow$ Tyrocidine-E ($1253.7Da$)

  by replacing the tyrosine residue at 7 with a phenylalanine residue ($-16Da$).

  Tyrocidine-A ($1269.7Da$) $\rightarrow$ Tyrocidine-B ($1308.7Da$)

  by replacing the phenylalanine residue at 3 with a tryptophan residue ($+39Da$).

- Tyrocidine-B ($1308.7Da$) $\rightarrow$ Tyrocidine-C($1347.7Da$)

  by replacing the D-phenylalanine residue at 4 with a D-tryptophan residue ($+39Da$).

- Tyrocidine-C ($1347.7Da$) $\rightarrow$ Tyrocidine-D($1370.7Da$)

  by replacing the tyrosine residue at 7 with a tryptophan residue ($+23Da$).

In this experiment, tyrocidine B, C, D and E were removed from the NRP database, leaving tyrocidine-A as the only tyrocidine. Therefore, B, C, D and E were considered analogs to A as if their structures were unknown. Analog search was executed iteratively after the initial database search, with the precursor mass window Δ set to $50Da$. If the algorithm works as expected, it should identify B and E in the first round using A as the seed. In the second round, by selecting B as the seed, C should be identified. If the first two rounds of analog search were successful, the third round should identify D with C as the seed. This progression is illustrated as arrows in Figure 4.2.

On top of that, as the LC-MS/MS data was previously examined in Experiment V, the database search results can be used to verify analog identifications. We use the tyrocidine identified in Experiment V as the reference result for an MS/MS scan. With analog search, if a scan is identified as an analog to tyrocidine-A and the reported structure is exactly the same as the reference, the analog identification would be concluded as *correct*. If the reported analog has the same total mass as the reference but the mass difference is located at a wrong site, the identification would be considered as *mis-localized*. We cannot judge the analog identifications on scans which were not identified in Experiment V, as there would be no reference in such case.

In the result, the initial database search identified tyrocidine-A from the LC-MS/MS dataset. The identification report is shown in Table 4.1. There were two identified NRPs,

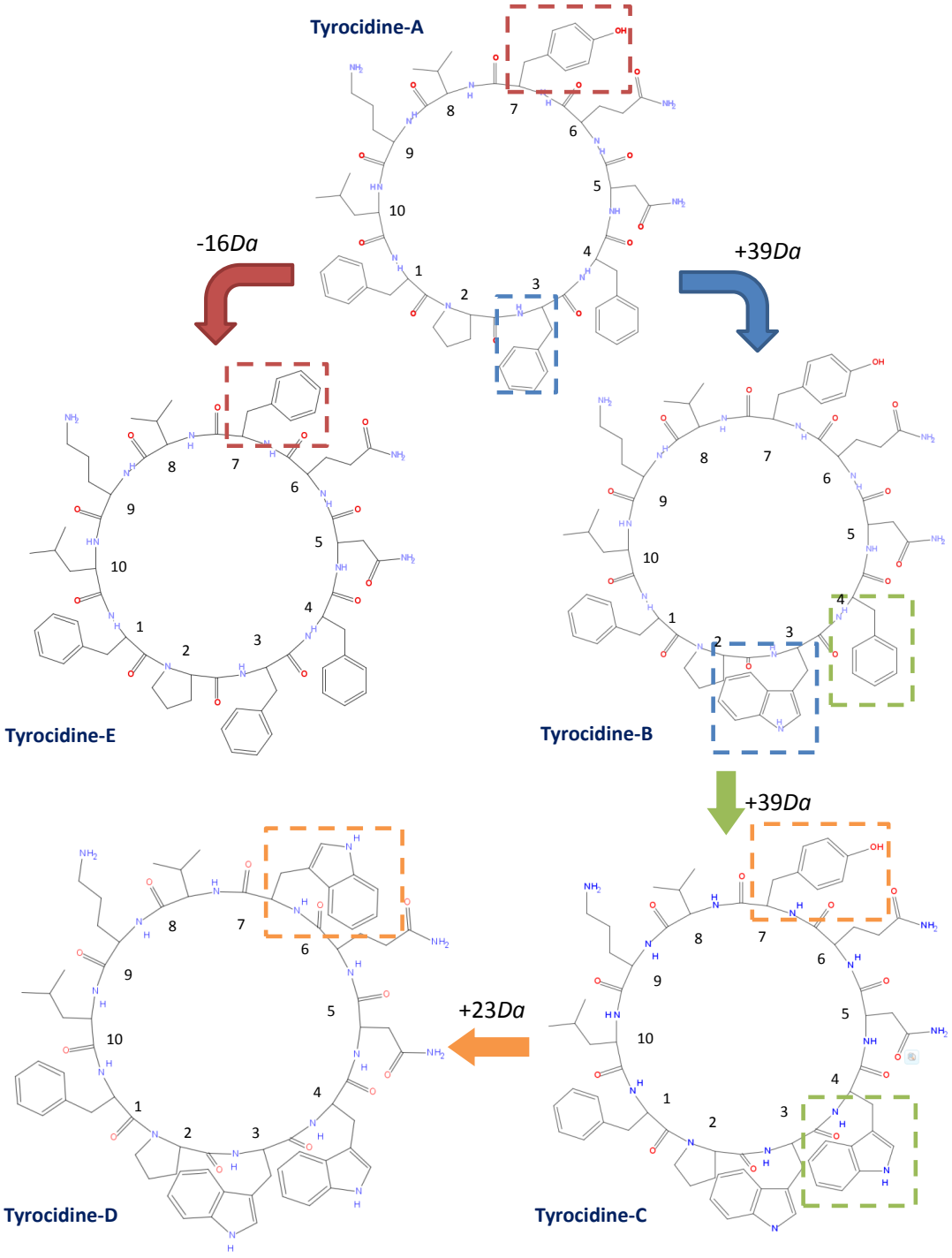


Figure 4.2: Structural comparison of tyrocidine A, B, C, D and E. Arrows in this figure indicate tyrocidine structures that only differ at one residue. Using A as the initial seed, analog search should identify B and E as analogs. If executed iteratively, analog search should identify B, C and D from A.

Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	Mass	P1 Score	P2 Score
1247	40.5	635.9	2	1269.8	Tyrocidine A	1269.65	92.8	43.3
1243	40.4	635.9	2	1269.8	Tyrocidine A	1269.65	89.9	42.5
1240	40.3	635.9	2	1269.8	Tyrocidine A	1269.65	89.9	66.8
1252	40.6	635.9	2	1269.7	Tyrocidine A	1269.65	67.4	42.1
1254	40.7	1271	1	1270	Tyrocidine A	1269.65	64.8	25.9
1258	40.8	635.7	2	1269.5	Tyrocidine A	1269.65	51.4	45.4
849	28.5	845.8	2	1689.7	Trichotoxin E	1689.02	32.5	33.6

Table 4.1: Identification report of the initial database search.

tyrocidine-A and trichotoxin-E. Tyrocidine-A was confidently identified with P_1=92.8, P_2=44.3. The P_1 score was higher than tyrocidine-A's P_1 score 84.7 in the previous database search experiment 3.4.6, because B, C, D, E were removed from the database. When there were no similar structures in database, the spectral match to tyrocidine-A became more significant. As for trichotoxin-E, we considered it as a false identification, because the scores were relatively low and it was only identified with one MS/MS scan.

The first round of analog search was performed with tyrocidine-A selected as the only seed. The identification report is shown in Table 4.2. In the table, MS/MS scans identified with analogs of approximately the same mass difference are grouped together, and highlighted by color shading. This reveals two scan groups among the 10 scans with analog identifications. The two groups, each consisting of 4 scans, account for the mass of tyrocidine B and E respectively. We checked the result in Experiment V, these scans were indeed identified as tyrocidine B and E previously. As such, iSNAP analog picked up the correct mass difference and identified tyrocidine B and E from the LC-MS/MS data. It successfully determined the location of the different building block on 3 out of 4 MS/MS scans in each group, and mis-localized the site for two scans in total.

In the second round, tyrocidine-B (tyrocidine-A+39.09@3) on scan 1162 was selected as the seed. The identification report is shown in Table 4.3. Among all identifications, tyrocidine-C was found as an analog to tyrocidine-B on three MS/MS scans. The difference site was localized correctly on scan 988 and 1137, and a wrong site was picked for scan

Analog search result									Comments	
Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	Modification	P_1 Score	P_2 Score	Reference result by database search	Conclusion
1270	41.1	1255	1	1254	Tyrocidine A_analog	-15.62@7	30.4	59.2	Tyrocidine E	correct
1259	40.8	627.9	2	1253.8	Tyrocidine A_analog	-15.86@7	29.8	37.7	Tyrocidine E	correct
1269	41.1	627.8	2	1253.7	Tyrocidine A_analog	-15.98@6	44.9	31.4	Tyrocidine E	mis-localized
1262	40.9	627.8	2	1253.7	Tyrocidine A_analog	-15.99@7	58.5	46.2	Tyrocidine E	correct
978	32.7	642.3	2	1282.7	Tyrocidine A_analog	+13.03@4	32.7	34.1	No result	unknown
1264	41	647.5	2	1292.9	Tyrocidine A_analog	+23.26@6	27.6	36.6	No result	unknown
1162	38.2	655.4	2	1308.7	Tyrocidine A_analog	+39.09@3	47	36.6	Tyrocidine B	correct
1190	38.9	655.4	2	1308.8	Tyrocidine A_analog	+39.15@9	27.3	38.8	Tyrocidine B	mis-localized
1172	38.5	655.5	2	1308.9	Tyrocidine A_analog	+39.25@3	27.7	49.4	Tyrocidine B	correct
1177	38.6	655.5	2	1308.9	Tyrocidine A_analog	+39.29@3	27.5	51.1	Tyrocidine B	correct

Table 4.2: Identification report of the first round of analog search using tyrocidine-A as the seed.

985. We select the highest scored tyrocidine-C (tyrocidine A+39.0@4+39.09@3) as seed for the third round (Table 4.4). Tyrocidine-D was identified on scan 1192, 1197 and 1202, and the site is correctly localized on scan 1197.

In the tables, we mark an identification as "unknown" if the MS/MS scan was not previously identified in Experiment V. Even though the analog identification is made with high P_1 and P_2 scores, we cannot conclude whether it is true. Such an analog identification can be one of other tyrocidines in the same family, or even for a novel analog. As stated before, there are more than 28 tyrocidines in the family, and the majority of them have mass around $1300Da$ [16]. To check these identifications, the general rule is to examine the number of identified scans in the same mass group. An analog identification is more likely to be true if more scans are identified as the same structure. Moreover, the mass difference need to be structurally explainable at the site. Finally, it is always worthwhile to manually check how well the spectrum is matched to the analog structure.

The experiments shows the iSNAP analog search has the basic capability in finding analog structures within a complex fermentation. It can find analogs differ from a seed NRP at one building block. However, iSNAP analog search is not fully automated. It requires human expertises in selecting seeds from identified structures and also in interpreting the search result. In this sense, iSNAP analog search is usable as an assistant tool for NRP analog deprelication and novel analog discovery. It exposes the MS/MS scans for possible

Analog search result									Comments	
Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	Modification	P_1 Score	P_2 Score	Reference result by database search	Conclusion
950	31.7	649.3	2	1296.6	Tyrocidine A_analog	-12.02@10, +39.09@3	27.9	24.2	No result	unknown
1264	41	647.5	2	1292.9	Tyrocidine A_analog	-15.75@6, +39.09@3	27.2	35.9	No result	unknown
978	32.7	642.3	2	1282.7	Tyrocidine A_analog	-25.98@10, +39.09@3	38.3	38	No result	unknown
929	31	637.8	2	1273.6	Tyrocidine A_analog	-35.07@10, +39.09@3	41.5	31.1	No result	unknown
967	32.3	661.8	2	1321.5	Tyrocidine A_analog	+12.86@7, +39.09@3	43.7	24.4	No result	unknown
1157	38	662.5	2	1322.9	Tyrocidine A_analog	+14.27@10, +39.09@3	31.1	39.5	No result	unknown
1164	38.2	662.5	2	1322.9	Tyrocidine A_analog	+14.27@9, +39.09@3	30	35.5	No result	unknown
1154	38	662.5	2	1323	Tyrocidine A_analog	+14.3@10, +39.09@3	31	37.7	No result	unknown
1035	34.6	663.4	2	1324.7	Tyrocidine A_analog	+16.06@10, +39.09@3	33.5	33.9	No result	unknown
1109	36.7	664.3	2	1326.5	Tyrocidine A_analog	+17.86@2, +39.09@3	56.3	24.1	No result	unknown
1257	40.8	666.9	2	1331.7	Tyrocidine A_analog	+23.02@10, +39.09@3	41	36.8	No result	unknown
1253	40.7	666.9	2	1331.7	Tyrocidine A_analog	+23.03@10, +39.09@3	42.4	33.3	No result	unknown
1249	40.5	666.9	2	1331.7	Tyrocidine A_analog	+23.07@10, +39.09@3	37.3	29.7	No result	unknown
1244	40.4	666.9	2	1331.9	Tyrocidine A_analog	+23.2@9, +39.09@3	27.5	33.4	No result	unknown
1102	36.5	667.8	2	1333.6	Tyrocidine A_analog	+24.89@8, +39.09@3	53.1	24.4	No result	unknown
942	31.5	668.8	2	1335.7	Tyrocidine A_analog	+26.99@10, +39.09@3	40.4	31.2	No result	unknown
947	31.6	668.8	2	1335.7	Tyrocidine A_analog	+27.0@10, +39.09@3	32.7	29.4	No result	unknown
937	31.3	668.9	2	1335.7	Tyrocidine A_analog	+27.08@10, +39.09@3	28.9	28.9	No result	unknown
922	30.8	668.9	2	1335.8	Tyrocidine A_analog	+27.09@7, +39.09@3	27.7	32.2	No result	unknown
1089	36.2	671.7	2	1341.4	Tyrocidine A_analog	+32.76@5, +39.09@3	29.3	24.9	No result	unknown
1225	39.9	673.8	2	1345.7	Tyrocidine A_analog	+37.0@10, +39.09@3	36.1	28.4	No result	unknown
1230	40	673.9	2	1345.7	Tyrocidine A_analog	+37.03@10, +39.09@3	27.6	28.5	No result	unknown
1220	39.8	673.9	2	1345.7	Tyrocidine A_analog	+37.05@10, +39.09@3	45.2	41.9	No result	unknown
1250	40.6	673.9	2	1345.7	Tyrocidine A_analog	+37.07@10, +39.09@3	27.7	29.8	No result	unknown
1255	40.7	673.9	2	1345.7	Tyrocidine A_analog	+37.07@8, +39.09@3	30.4	36.3	No result	unknown
988	33	674.8	2	1347.7	Tyrocidine A_analog	+39.0@4, +39.09@3	32.1	29.6	Tyrocidine C	correct
985	32.9	674.9	2	1347.7	Tyrocidine A_analog	+39.05@9, +39.09@3	27.6	28.7	Tyrocidine C	mis-localized
1137	37.5	674.9	2	1347.9	Tyrocidine A_analog	+39.22@4, +39.09@3	28.4	33.6	Tyrocidine C	correct

Table 4.3: Identification report of the second round of analog search using tyrocidine-B (A+39.25@3) as the seed.

Analog search result									Comments	
Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	Modification	P_1 Score	P_2 Score	Reference result by database search	Conclusion
947	31.6	668.8	2	1335.7	Tyrocidine A_analog	-12.01@10, +39.0@4, +39.09@3	35.5	31	No result	unknown
942	31.5	668.8	2	1335.7	Tyrocidine A_analog	-12.02@10, +39.0@4, +39.09@3	41.4	31.3	No result	unknown
952	31.8	668.8	2	1335.6	Tyrocidine A_analog	-12.11@7, +39.0@4, +39.09@3	44.7	24	No result	unknown
1102	36.5	667.8	2	1333.6	Tyrocidine A_analog	-14.12@7, +39.0@4, +39.09@3	55.2	25.6	No result	unknown
1088	36.1	1333.2	1	1332.2	Tyrocidine A_analog	-15.44@1, +39.0@4, +39.09@3	29.1	31	No result	unknown
1082	36	667	2	1332	Tyrocidine A_analog	-15.64@10, +39.0@4, +39.09@3	27.3	25.8	No result	unknown
1244	40.4	666.9	2	1331.9	Tyrocidine A_analog	-15.81@7, +39.0@4, +39.09@3	34.9	46.5	No result	unknown
1249	40.5	666.9	2	1331.7	Tyrocidine A_analog	-15.94@10, +39.0@4, +39.09@3	38.7	30.8	No result	unknown
1253	40.7	666.9	2	1331.7	Tyrocidine A_analog	-15.98@10, +39.0@4, +39.09@3	41.9	32.8	No result	unknown
1257	40.8	666.9	2	1331.7	Tyrocidine A_analog	-15.99@10, +39.0@4, +39.09@3	37.7	33.9	No result	unknown
1035	34.6	663.4	2	1324.7	Tyrocidine A_analog	-22.95@10, +39.0@4, +39.09@3	30	29.9	No result	unknown
1275	41.3	658.3	2	1314.5	Tyrocidine A_analog	-33.13@10, +39.0@4, +39.09@3	56.6	26.4	No result	unknown
912	30.5	657.4	2	1312.7	Tyrocidine A_analog	-34.95@10, +39.0@4, +39.09@3	29.9	31.1	No result	unknown
917	30.6	657.4	2	1312.7	Tyrocidine A_analog	-34.98@10, +39.0@4, +39.09@3	28.6	27.7	No result	unknown
924	30.9	657.3	2	1312.6	Tyrocidine A_analog	-35.1@3, +39.0@4, +39.09@3	42.1	24.3	No result	unknown
1150	37.8	681.3	2	1360.7	Tyrocidine A_analog	+12.97@10, +39.0@4, +39.09@3	54.4	29.8	No result	unknown
1124	37.1	681.9	2	1361.8	Tyrocidine A_analog	+14.14@10, +39.0@4, +39.09@3	34.6	40	No result	unknown
1118	37	681.9	2	1361.9	Tyrocidine A_analog	+14.2@9, +39.0@4, +39.09@3	31.7	38.4	No result	unknown
1113	36.8	682	2	1361.9	Tyrocidine A_analog	+14.26@10, +39.0@4, +39.09@3	31.1	37	No result	unknown
1054	35.1	682.4	2	1362.9	Tyrocidine A_analog	+15.18@8, +39.0@4, +39.09@3	28.8	40.6	No result	unknown
968	32.3	682.8	2	1363.6	Tyrocidine A_analog	+15.97@10, +39.0@4, +39.09@3	42.3	29.8	No result	unknown
982	32.8	682.9	2	1363.7	Tyrocidine A_analog	+16.04@10, +39.0@4, +39.09@3	37.8	36.9	No result	unknown
977	32.6	682.9	2	1363.9	Tyrocidine A_analog	+16.2@10, +39.0@4, +39.09@3	33.7	42.4	No result	unknown
1083	36	683.4	2	1364.7	Tyrocidine A_analog	+17.03@10, +39.0@4, +39.09@3	33.2	31.3	No result	unknown
1070	35.6	683.7	2	1365.4	Tyrocidine A_analog	+17.74@10, +39.0@4, +39.09@3	58	28.9	No result	unknown
1072	35.7	683.8	2	1365.6	Tyrocidine A_analog	+17.97@10, +39.0@4, +39.09@3	61.5	33.6	No result	unknown
1202	39.3	686.4	2	1370.8	Tyrocidine A_analog	+23.13@8, +39.0@4, +39.09@3	31.5	39	Tyrocidine D	mis-localized
1197	39.1	686.4	2	1370.9	Tyrocidine A_analog	+23.19@7, +39.0@4, +39.09@3	31.3	45.2	Tyrocidine D	correct
1192	39	686.5	2	1370.9	Tyrocidine A_analog	+23.22@8, +39.0@4, +39.09@3	30.6	40.1	Tyrocidine D	mis-localized
1149	37.8	688.4	2	1374.7	Tyrocidine A_analog	+27.02@10, +39.0@4, +39.09@3	38.2	32.5	No result	unknown
923	30.8	688.4	2	1374.7	Tyrocidine A_analog	+27.04@1, +39.0@4, +39.09@3	27.2	24.2	No result	unknown
997	33.3	694.3	2	1386.6	Tyrocidine A_analog	+38.94@10, +39.0@4, +39.09@3	50.2	27.5	No result	unknown
992	33.2	694.3	2	1386.6	Tyrocidine A_analog	+38.95@10, +39.0@4, +39.09@3	52.4	26.8	No result	unknown
1012	33.8	694.3	2	1386.7	Tyrocidine A_analog	+39.01@10, +39.0@4, +39.09@3	44.8	36.7	No result	unknown
1007	33.7	694.4	2	1386.7	Tyrocidine A_analog	+39.02@10, +39.0@4, +39.09@3	46.3	30.9	No result	unknown
1002	33.5	694.4	2	1386.7	Tyrocidine A_analog	+39.04@10, +39.0@4, +39.09@3	29.3	29.6	No result	unknown
1120	37	694.8	2	1387.7	Tyrocidine A_analog	+39.98@10, +39.0@4, +39.09@3	38.5	28.4	No result	unknown

Table 4.4: Identification report of the third round of analog search using tyrocidine-C (A+39.04@4+39.25@3) as the seed.

analogs and also the analogs' structural information. A conclusive characterization of the analog structure needs further experimental techniques (such as NMR). However, the software's result provide a list of good candidates for the follow-up experiments to start with.

Chapter 5

Future work

5.1 Detection of NRP Analogs with Specified Modifications using Database Search

Novel NRPs can be found as analogs to a known NRP. They can have nearly identical structure but only differ at a few monomers. A database search algorithm would be more useful in NRP discovery if it can identify the analogs of known NRPs in the database. This is the very incentive for creating the iSNAP analog search algorithm described in Chapter 4. As a prototype, it fulfills the objectives, but has two major drawbacks in practical use. First, in order to have analogs identified with analog search, it requires the NRP in the original form be identified in the initial database search. As such, analogs search is nullified if the original form does not exist in the analyzed sample at all. Secondly, the number of modified building blocks is limited to one in analog search. Though the analog search algorithm can be used iteratively, it cannot find analogs with more than one difference sites, if there is no analog with exactly one of the difference sites to intermediate the progression. As an example, in Experiment VI, if tyrocidine-C is not in the sample, it is not possible for the iterative analog search to find tyrocidine-D as an analog to tyrocidine-A.

The benefit of the above analog search is that we do not have to know the mass difference on the modified site (Δm) prior to the search. If we have the knowledge of what the possible changes are, a different search strategy can be designed. With the input of a small set of $\{\Delta m_i\}$ $(i = 1 \ldots k)$, as well as the maximum number l $(l < k)$ of sites that can be modified on a database NRP, the problem becomes analogous to the traditional database search with specified post-translational modifications[44]. The

problem can be solved similarly. For an input tandem spectrum with precursor mass of M, we can enumerate subcollections of $\{\Delta m_i\}$ with a maximum of l elements. For each subcollection $\{\Delta \hat{m}_j\}$ $(1 \leq j \leq \hat{l} \leq l, \Delta \hat{m}_j \in \{\Delta m_i\})$, the mass differences $\Delta \hat{m}_j$ $(1 \leq j \leq \hat{l})$ are exhaustively apply to database NRPs with mass around $M - \sum_{j=1}^{\hat{l}} \Delta \hat{m}_j$, so that analog candidates can be generated. These candidates can be scored using the same scoring scheme for database search.

The method is computationally intensive and definitely needs further work. However, it does not require the original form to be in the sample and can possibly support analog identification with a few more modified sites.

5.2 Nonribosomal Peptide Identification with Spectra Library

Spectral library search is one of the approaches for peptide identification. It is considered to have better identification accuracy and can be faster than searching a structural database. Assuming we have a library consisting of previously generated tandem spectra of nonribosomal peptides, the NRP identification problem could be solved much easier. With such a spectral library, the structural complexity of NRP is no longer an issue, all the algorithm need to do is to match the input spectra with the spectra in the library. Such a library search algorithm for NRPs is intrinsically the same as the library search algorithm for linear peptides.

Specifically, using an NRP spectra can be beneficial in at least two aspects. The first is, fragment ions generated from unexpected pathways can now be matched, and contributed to a more reliable identification. It is specially important for NRP tandem spectra with the presence of non-directed sequences (NDS). These ions are supposed to appear in both the input spectra and in the library spectrum of the corresponding NRP, thus would no longer be "unexpected". Secondly, library spectra have the ion intensity information as well as the m/z value of each fragment. Hypothetical fragments used in structural database search only have the m/z. When a fragment ion with high intensity is matched, it should be considered more significant than a match of low intensity ions. Ion intensities, as extra information can help evaluate the matching quality and ultimately lead to more reliable scores.

However, NRP spectral library search is hindered by the availability of such a library. Above all, there is no compiled spectral library with a considerable number of nonribosomal peptides. If we have partners in biochemical laboratories willing to collaborate on

this project, building such a library would have great impact not only for bioinformatics research, but also for the research of novel NRP discovery. The challenge is, there is no ready supply of the nonribosomal peptides. Only a small number of NRPs can be ordered from commercial suppliers. For the others, microbial fermentation will need to be done, and the NRP need to be screened and exacted. It will take tremendous amount of work. Moreover, a protocol need to be designed for compiling the library, which regulates the procedure of running mass spectrometers and also the pre-process of mass spectra.

5.3 Targeted NRP Identification by Searching Database of Predicted NRPs

As introduced in Chapter 2, it is understood that nonribosomal peptide synthetases (NRPS), occasionally as well as polyketide synthases (PKS), are involved in the biosynthesis of nonribosomal peptides. NRPS and PKS are encoded in gene clusters within microbial genomes. These synthetases are the modules that assemble NRPs and lead to the diversity we have seen[14].

As bioinformatics research advances, there are already software that can identify gene clusters encoding NRPS and PKS by analyzing the genome sequence. AntiSMASH[45], for instance, claimed to enable genome-wide rapid identification, annotation and analysis of gene clusters for secondary metabolite biosynthesis. With such technology, NRPS and PKS for the biosynthesis of secondary metabolites can be revealed. NRPS and PKS modules can be translated to building blocks in the secondary metabolite, and allow researchers to predict the molecular structure of novel NRPs.

Our research collaborators in Nathan Magarvey Lab at McMaster University perceives a genome-wide analysis using AntiSMASH would facilitate the prediction of many novel NRPs, which may be produced by the bacteria. These predicted NRPs could be added into the build-in database, and then microbial fermentations cultured in varying conditions can be screened by iSNAP. With this targeted approach, iSNAP could directly contribute to the discover of novel NRPs.

Chapter 6

Summary

In this work, iSNAP has been proposed as the first database search software for non-ribosomal peptides. It has been evolved from the traditional database search approach to be competent at handling complicated NRP structures. NRP identifications are made by comparing the input MS/MS spectra with hypothetical fragments generated from NRPs in the database. Significance scores P_1 and P_2 are calculated to indicate the confidence of an NRP identification. Combined with an in-house structural database of 1107 NRPs, iSNAP was tested to be an effective tool in revealing the true NRP for input MS/MS scans. Experiments demonstrated that the database search algorithm could find spiked and naturally produced NRPs within complex samples, and could distinctively identify mixed NRPs with similar structures. Across identified MS/MS scans, the false discovery rate was tested to be less than 7%. As such, iSNAP enables automated and high-throughput NRP dereplication and can faciliate the discovery of novel NRPs.

Furthermore, iSNAP analog search has been developed as an extension to the NRP database search algorithm. It is a semi-automated software to identify NRP analogs, which are outside of the NRP database, with one building block differ from a previouly identified NRP. iSNAP analog search is usable as an assistant tool for NRP analog deprelication and novel analog discovery. It exposes a list of MS/MS scans for potential analogs and also the analogs' structural information. iSNAP provides a list of good candidates for follow-up experiments to start with, and helps move towards the conclusive characterization of novel NRP analogs.

The iSNAP as a web service is available at http://monod.uwaterloo.ca/isnap.

Appendix A

Appendix

A.1 iSNAP Web Service

A.1.1 User Interface of iSNAP

See Figure A.1 A.2

A.1.2 How to Use iSNAP

- Step 1: Convert mass spectral data to .mzXML
  - iSNAP supports .mzXML as input, which is a standardized format for mass spectral data. Instrument vendors usually provide free software that convert native acquisitions to this standard format. For instance, ReAdW can be used to convert ThermoFinnigan raw files, and CompassXport for Bruker raw files, etc. Besides, there are also third-party efforts trying to simplify the conversion. ProteoWizard's msconvert supports the conversion of Agient, Bruker, Thermo, Waters and AB Sciex file formats into mzXML.
- Step 2: Input an .mzXML file
  - Click "Choose File", and select the .mzXML file with the pop-up dialog. We provided a test example, the LC-MS/MS data of *Bacillus sp.* fermentation extract. The example data can be download by clicking the link "Example .mzXML file".

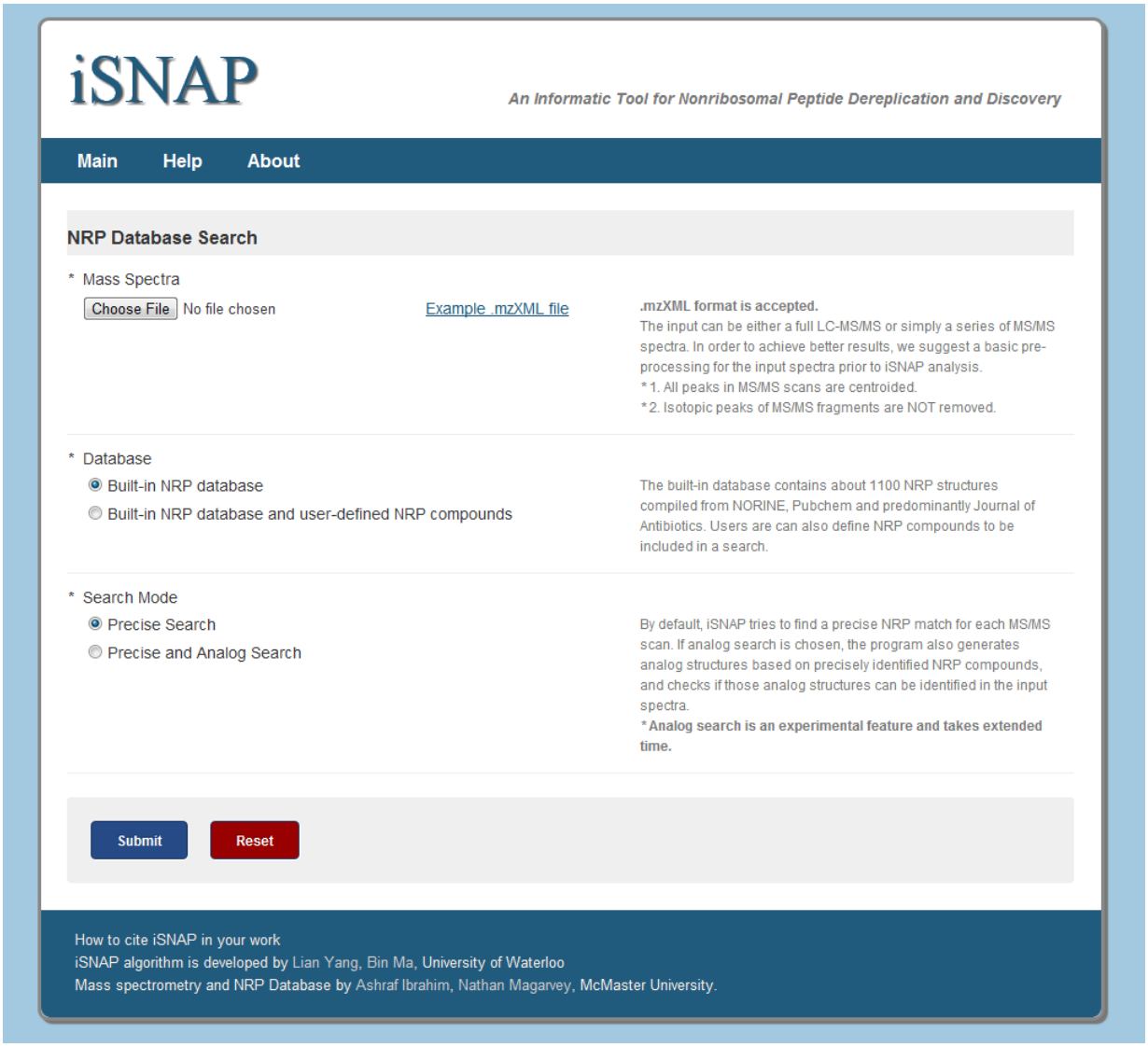


Figure A.1: User interface of iSNAP web service.

Summary

Upload — Done.

Database Search — Done.

Reports — Complete NRP Search Report
Download

Click labels to sort.

Scan No.	RT	Precursor m/z	Precursor charge	Precursor mass	Result	SMILES	Mass	P1 Score	P2 Score
1247	40.5	636.2	2	1270.4	Tyrocidine A	CC(C)CC1NC(=O)C(CCCN)NC(=O)C(NC(=O)C(CC2=CC=C(O)C=C2)NC(=O)C(CCC(N)=O)NC(=O)C(CC(N)=O)NC(=O)C(CC3=CC=CC=C3)NC(=O)C(CC4=CC=CC=C4)NC(=O)C5CCCN5C(=O)C(CC6=CC=CC=C6)NC1=O)C(C)C	1269.65	92.8	43.3
1243	40.4	636.2	2	1270.4	Tyrocidine A	CC(C)CC1NC(=O)C(CCCN)NC(=O)C(NC(=O)C(CC2=CC=C(O)C=C2)NC(=O)C(CCC(N)=O)NC(=O)C(CC(N)=O)NC(=O)C(CC3=CC=CC=C3)NC(=O)C(CC4=CC=CC=C4)NC(=O)C5CCCN5C(=O)C(CC6=CC=CC=C6)NC1=O)C(C)C	1269.65	89.9	42.5
1240	40.3	636.2	2	1270.4	Tyrocidine A	CC(C)CC1NC(=O)C(CCCN)NC(=O)C(NC(=O)C(CC2=CC=C(O)C=C2)NC(=O)C(CCC(N)=O)NC(=O)C(CC(N)=O)NC(=O)C(CC3=CC=CC=C3)NC(=O)C(CC4=CC=CC=C4)NC(=O)C5CCCN5C(=O)C(CC6=CC=CC=C6)NC1=O)C(C)C	1269.65	89.9	66.8
1252	40.6	636.2	2	1270.3	Tyrocidine A	CC(C)CC1NC(=O)C(CCCN)NC(=O)C(NC(=O)C(CC2=CC=C(O)C=C2)NC(=O)C(CCC(N)=O)NC(=O)C(CC(N)=O)NC(=O)C(CC3=CC=CC=C3)NC(=O)C(CC4=CC=CC=C4)NC(=O)C5CCCN5C(=O)C(CC6=CC=CC=C6)NC1=O)C(C)C	1269.65	67.4	42.1
1254	40.7	1271.3	1	1270.3	Tyrocidine A	CC(C)CC1NC(=O)C(CCCN)NC(=O)C(NC(=O)C(CC2=CC=C(O)C=C2)NC(=O)C(CCC(N)=O)NC(=O)C(CC(N)=O)NC(=O)C(CC3=CC=CC=C3)NC(=O)C(CC4=CC=CC=C4)NC(=O)C5CCCN5C(=O)C(CC6=CC=CC=C6)NC1=O)C(C)C	1269.65	64.8	25.9
1258	40.8	636.0	2	1270.1	Tyrocidine A	CC(C)CC1NC(=O)C(CCCN)NC(=O)C(NC(=O)C(CC2=CC=C(O)C=C2)NC(=O)C(CCC(N)=O)NC(=O)C(CC(N)=O)NC(=O)C(CC3=CC=CC=C3)NC(=O)C(CC4=CC=CC=C4)NC(=O)C5CCCN5C(=O)C(CC6=CC=CC=C6)NC1=O)C(C)C	1269.65	51.4	45.4
1172	38.5	655.8	2	1309.5	Tyrocidine A analog	O=C(N)CCC1NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C4N(C(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)[+39.85C](NC1(=O))CC2=CC=C(O)C=C2)C(C)C)CCCN)CC(C)C)CC3=CC=CC=C3)CCC4)CC5=CC=CC=C5)CC6=CC=CC=C6)CC(=O)N	1269.65 +39.85	51.1	26.3
1177	38.6	655.8	2	1309.5	Tyrocidine A analog	O=C(N)CC[+39.83C]1NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C4N(C(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC1(=O))CC2=CC=C(O)C=C2)C(C)C)CCCN)CC(C)C)CC3=CC=CC=C3)CCC4)CC5=CC=CC=C5)CC6=CC=CC=C6)CC(=O)N	1269.65 +39.83	48.9	26.2

Figure A.2: A sample identification report generated by iSNAP.

- Step 3: Choose the database
  - By default, iSNAP search with the built-in database which was assembled in-house, contains about 1107 nonribosomal peptides. Users can also upload their own NRPs to be included into the database for a search. These NRPs should be encoded as SMILES code and formatted in text file with each NRP takes one line.
- Step 4: Specify the search mode
  - "Precise Search" is the default search mode. In this mode, database search is perform with algorithm described in Chapter 3. Identifications can only be made when the MS/MS spectrum matches an NRP in the database. In the "Precise and Analog Search" mode, the analog search feature turned on. It runs the analog search described in Chapter 4 after the initial "Precise Search" is done. The size of analog search window Δ need to be specified for analog search.
- Step 5: Submit the task
  - By simply clicking the Submit button, iSNAP will start to analyze the input spectra and perform NRP identification. The web page should not be closed throughout the analyzing process. The progress will be updated on the web, until the analysis is finished.
- Step 6: Understand the identification report
  - An identification report will be summarized for your inspection after iSNAP finishes the analysis. The report is formatted in Excel spreadsheet and can be downloaded with a link. For each MS/MS scan, NRP candidates in $\pm 1 Da$ precursor mass window are listed in the report with their information and scores. Candidates are sorted by P_1 score. Both P_1 and P_2 scores indicate confidence of the identification, which is explained in Chapter 3. Besides the report, a brief summary will be displayed on the web page. The summary only shows the identified NRPs, which are the top candidates corresponding the MS/MS spectrum and have P_1 and P_2 above score thresholds.

A.1.3 Technical Details

The iSNAP web service is developed as a Java Servlet, deployed on the Apache web server hosted on the www-novo.cs.uwaterloo.ca. The program is designed to support simultaneous sessions from multiple users. It dispatches search tasks by assigning a dedicated thread for each. In the stage of analog search, MS/MS spectra are divided into batches and analyzed with multiple threads, making use of all available CPUs, as this step is computationally intensive.

A.1.4 Acknowledgment

iSNAP uses following software packages as libraries.

- The Chemistry Development Kit (CDK)[46], under LGPL license.
- Java Excel API (JExcelApi), under LGPL license.

The build-in NRP database is compiled by Ashraf Ibrahim and Nathan Magarvey from McMaster University.

References

[1] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the last 25 years. *Journal of Natural Products*, 70(3):461–77, March 2007.

[2] Mark S Butler. The role of natural product chemistry in drug discovery. *Journal of Natural Products*, 67(12):2141–53, December 2004.

[3] Luciana S Cardoso, Maria Ilma Araujo, Alfredo M Góes, Lucila G Pacífico, Ricardo R Oliveira, and Sergio C Oliveira. Polymyxin B as inhibitor of LPS contamination of Schistosoma mansoni recombinant proteins in human cytokine analysis. *Microbial Cell Factories*, 6(1):1, January 2007.

[4] Jesse W-H Li and John C Vederas. Drug discovery and natural products: end of an era or an endless frontier? *Science*, 325:161–165, July 2009.

[5] A MA van Wageningen, PN Kirkpatrick, DH Williams, and BR Harris. Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chemistry & Biology*, 5(3):155–162, 1998.

[6] Peter Kirkpatrick, Aarti Raja, Jason LaBonte, and John Lebbos. Daptomycin. *Nature Reviews. Drug Discovery*, 2(12):943–944, December 2003.

[7] Kazuko Hori, Yoshihiro Yamamoto, Toshitsugu Kurotsu, Masayuki Kanda, Setsuko Miura, Kaori Okamura, Junichi Furuyama, and Yoshitaka Saito. Molecular cloning and nucleotide sequence of the gramicidin S synthetase 1 gene. *Journal of Biochemistry*, 106(4):639–645, 1989.

[8] D.R. Storm, K.S. Rosenthal, and P.E. Swanson. Polymyxin and related peptide antibiotics. *Annual Review of Biochemistry*, 46(1):723–763, 1977.

[9] H F Stähelint. The history of cyclosporin A (Sandimmune) revisited: another point of view. *Experientia*, 52(1):5–13, 1996.

[10] Cllaud Vezin, Alicia Kudelski, and S N Sehgal. Rapamycin (AY-22,989), a new antifungal antibiotic. *The Journal of Antibiotics*, 28(10):721–726, 1975.

[11] H Umezawa, K Maeda, T Takeuchi, and Y Okami. New antibiotics, bleomycin A and B. *Journal of Antibiotics (Tokyo)*, 19(5):200–209, 1966.

[12] I Molnár, T Schupp, M Ono, R Zirkle, M Milnamow, B Nowak-Thompson, N Engel, C Toupet, A Stratmann, D D Cyr, J Gorlach, J M Mayo, A Hu, S Goff, J Schmid, and J M Ligon. The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from Sorangium cellulosum So ce90. *Chemistry & Biology*, 7(2):97–109, February 2000.

[13] Ségolène Caboche, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov. NORINE: a database of nonribosomal peptides. *Nucleic Acids Research*, 36(Database issue):D326–31, January 2008.

[14] James B McAlpine. Advances in the understanding and use of the genomic base of microbial secondary metabolite biosynthesis for the discovery of new natural products. *Journal of natural products*, 72(3):566–72, March 2009.

[15] Dudley H Williams and Ian Fleming. *Spectroscopic methods in organic chemistry*, volume 4. McGraw-Hill, 1995.

[16] XJ Tang, Pierre Thibault, Robert K Boyd, Elsevier Science Publishers B V, and K Boyd. Characterisation of the tyrocidine and gramicidin fractions of the tyrothricin complex from Bacillus brevis using liquid chromatography and mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes*, 122(3487):153–179, 1992.

[17] Robert Finking and Mohamed a Marahiel. Biosynthesis of nonribosomal peptides. *Annual Review of Microbiology*, 58:453–88, January 2004.

[18] Dirk Schwarzer, Robert Finking, and Mohamed a. Marahiel. Nonribosomal peptides: from genes to products. *Natural Product Reports*, 20(3):275, 2003.

[19] W Gevers, H Kleinkauf, and F Lipmann. The activation of amino acids for biosynthesis of gramicidin S*. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 269–276, 1968.

[20] CT Walsh, Huawei Chen, TA Keating, and BK Hubbard. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Current Opinion in Chemical Biology*, 5(5):525–534, 2001.

[21] Michael a Fischbach and Christopher T Walsh. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chemical Reviews*, 106(8):3468–96, August 2006.

[22] Jimmy K Eng, Ashley L Mccormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[23] David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and Jogn S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, 1999.

[24] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles a Lajoie, and Bin Ma. PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics*, 11(4), December 2011.

[25] Robertson Craig and Ronald C Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–7, June 2004.

[26] Mitchell J. Wells and S A McLuckey. Collision-induced dissociation (CID) of peptides and proteins. *Methods in Enzymology Biological Mass Spectrometry*, 402:148–185, 2005.

[27] Béla Paizs and Sándor Suhai. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24(4):508–548, 2005.

[28] Hanno Steen and Matthias Mann. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews. Molecular Cell Biology*, 5(9):699–711, September 2004.

[29] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–42, January 2003.

[30] J a Taylor and R S Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, 73(11):2594–604, June 2001.

[31] Ari Frank and Pavel Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4):964–73, February 2005.

[32] Frank Desiere, Eric W Deutsch, Alexey I Nesvizhskii, Parag Mallick, Nichole L King, Jimmy K Eng, Alan Aderem, Rose Boyle, Erich Brunner, Samuel Donohoe, Nelson Fausto, Ernst Hafen, Lee Hood, Michael G Katze, Kathleen a Kennedy, Floyd Kregenow, Hookeun Lee, Biaoyang Lin, Dan Martin, Jeffrey a Ranish, David J Rawlings, Lawrence E Samelson, Yuzuru Shiio, Julian D Watts, Bernd Wollscheid, Michael E Wright, Wei Yan, Lihong Yang, Eugene C Yi, Hui Zhang, and Ruedi Aebersold. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology*, 6(1):R9, January 2005.

[33] K Eckart. Mass spectrometry of cyclic peptides. *Mass Spectrometry Reviews*, 13:23–55, 1994.

[34] Wei-Ting Liu, Julio Ng, Dario Meluzzi, Nuno Bandeira, Marcelino Gutierrez, Thomas L Simmons, Andrew W Schultz, Roger G Linington, Bradley S Moore, William H Gerwick, Pavel a Pevzner, and Pieter C Dorrestein. Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. *Analytical Chemistry*, 81(11):4200–9, June 2009.

[35] Alex G Harrison, Alex B Young, Christian Bleiholder, Sandor Suhai, and Béla Paizs. Scrambling of sequence information in collision-induced dissociation of peptides. *Journal of the American Chemical Society*, 128(32):10364–5, August 2006.

[36] Julio Ng, Nuno Bandeira, W.T. Liu, Majid Ghassemian, T.L. Simmons, W.H. Gerwick, Roger Linington, P.C. Dorrestein, and P.A. Pevzner. Dereplication and de novo sequencing of nonribosomal peptides. *Nature Methods*, 6(8):596–599, 2009.

[37] P.A. PA Pevzner, V. Dancik, and C.L. Tang. Mutation-tolerant protein identification by mass spectrometry. *Journal of Computational Biology*, 7(6):777–787, 2000.

[38] D Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[39] Anders Broberg, Audrius Menkis, and Rimvydas Vasiliauskas. Kutznerides 1-4, depsipeptides from the actinomycete Kutzneria sp. 744 inhabiting mycorrhizal roots of Picea abies seedlings. *Journal of Natural Products*, 69(1):97–102, January 2006.

[40] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7(1):29–34, January 2008.

[41] Anton Pohanka, Audrius Menkis, Jolanta Levenfors, and Anders Broberg. Low-abundance kutznerides from Kutzneria sp. 744. *Journal of Natural Products*, 69(12):1776–81, 2006.

[42] Michael J MacCoss, Christine C Wu, and John R Yates. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical chemistry*, 74(21):5593–9, November 2002.

[43] David M Creasy and John S Cottrell. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*, 2(10):1426–34, October 2002.

[44] Stephen Tanner, Hongjun Shu, Ari Frank, and LC Wang. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*, 77(14):4626–39, 2005.

[45] Marnix H Medema, Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael a Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, 39(Web Server issue):W339–46, July 2011.

[46] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003.